Leveraging Subword Embeddings for Multinational Address Parsing

Marouane Yassine, David Beauchemin,
François Laviolette, Luc Lamontagne

Département d'informatique et de génie logiciel,
Université Laval

*marouane.yassine.1@ulaval.ca, david.beauchemin.5@ulaval.ca,
francois.laviolette@ift.ulaval.ca, luc.lamontagne@ift.ulaval.ca*

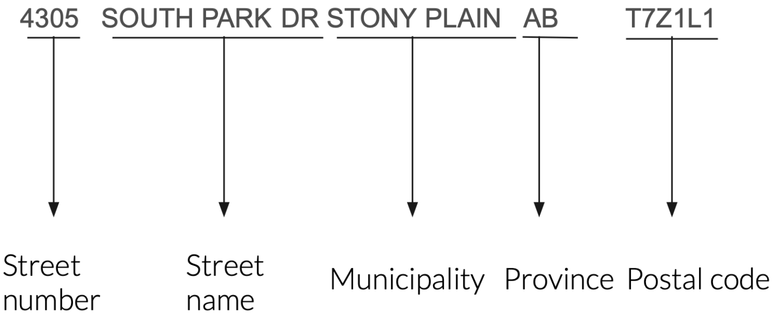
July 14 2020

Groupe de
Recherche en
Apprentissage
Automatique de
Laval

UNIVERSITÉ
LAVAL

Agenda

What's address parsing ?



Useful for tasks such as *Record Linkage* and *Geocoding*

Agenda

Related Work - Address parsing for one country

- Rule based methods
- Probabilistic models
  - Hidden Markov Models (HMM)
  - Conditional Random Fields (CRF)
- Neural networks
  - Feed-forward Neural Network
  - Recurrent Neural Networks

Related work - Multinational address parsing

Libpostal[1]

- CRF based model
- Preprocessing
- '*trained on over 1 billion examples in every inhabited country on Earth*'[2]

No previous neural network approaches for multinational address parsing

1. https://github.com/openvenues/libpostal
2. https://medium.com/@albarrentine/statistical-nlp-on-openstreetmap-part-2-80405b988718

Agenda

1 Introduction

2 Related Work
- Address parsing for one country
- Multinational address parsing

3 Subword Embeddings

4 Architecture

5 Data

6 Experiments

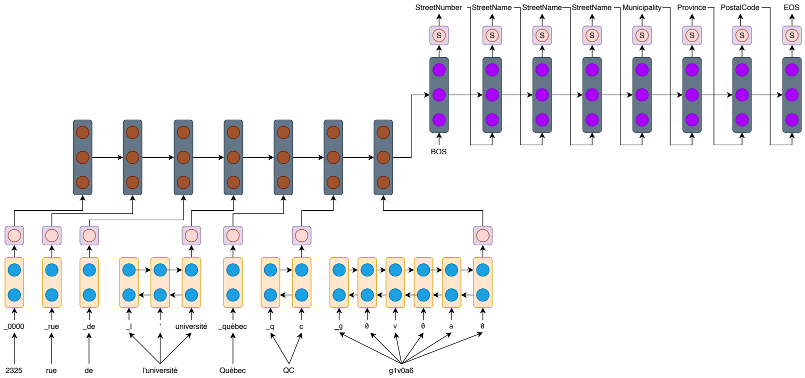
7 Conclusion

Subword Embeddings

...

Agenda

Architecture



Agenda

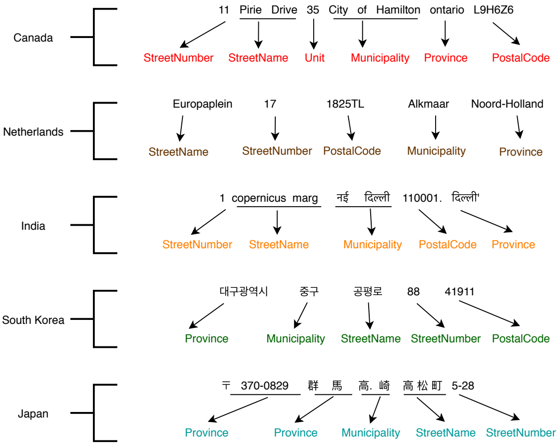
About the Data

- Built using the open-source data on which Libpostal's models were trained.
- Contain 61 countries.
- We used eight tags : StreetNumber, StreetName, Unit, Municipality, Province, PostalCode, Orientation, and GeneralDeliver[3].

3. Libpostal used 20 tags.

Examples of Address and their patterns

We use five different address patterns (one for each color) and a sixth one for some countries' using more than a pattern (no color).



Training Data and Holdout Test Set

20 countries are used for the multinational training with a sample size of 100,000 per country and the rest is used as a holdout.

Country	Number of samples	Country	Number of samples	Country	Number of samples	Country	Number of samples
United States	8,000,000	Germany	1,576,059	Poland	459,522	Czechia	195,269
Brazil	8,000,000	Spain	1,395,758	Norway	405,649	Italy	178,848
South Korea	6,048,106	Netherlands	1,202,173	Austria	335,800	France	20,050
Australia	5,428,043	Canada	910,891	Finland	280,219	United Kingdom	14,338
Mexico	4,853,349	Switzerland	474,240	Denmark	199,694	Russia	8115

Zero-Shot Test Set

41 countries are used for zero-shot transfer evaluation (never seen in training).

Country	Number of samples	Country	Number of samples	Country	Number of samples	Country	Number of samples
Belgium	66,182	Slovenia	9773	Réunion	2514	Singapore	968
Sweden	32,291	Ukraine	9554	Moldova	2376	Bangladesh	888
Argentina	27,692	Belarus	7590	Indonesia	2259	Paraguay	839
India	26,084	Serbia	6792	Bermuda	2065	Cyprus	836
Romania	19,420	Croatia	5671	Malaysia	2043	Bosnia	681
Slovakia	18,975	Greece	4974	South Africa	1388	Ireland	638
Hungary	17,460	New Zealand	4678	Latvia	1325	Algeria	601
Japan	14,089	Portugal	4637	Kazakhstan	1087	Colombia	569
Venezuela	10,696	Lithuania	3126	New Caledonia	1036	Uzbekistan	505
Philippines	10,471	Faroe Islands	2982	Estonia	1024		

Agenda

Training Procedure

- We trained five times[4] each of our two model (**fastText** and **BPEmb**).
- 200 epochs with a batch size of 2048.
- We used an early stopping with a patience of 15 epochs.
- A starting learning rate at 0.1 and a learning rate scheduling (factor of 0.1) after ten epochs without loss decrease.
- Cross-Entropy loss.
- Stochastic Gradient Descent (SGD) optimizer.
- Teacher forcing [**?**]

4. With the following seed $\{5, 10, 15, 20, 25\}$

Multinational Evaluation

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

Zero-shot Evaluation

Country	FastText	BPEmb	Country	FastText	BPEmb
Belgium	**88.14 ± 1.04**	87.45 ± 1.37	Faroe Islands	74.14 ± 1.83	**86.59 ± 2.21**
Sweden	81.59 ± 4.53	**88.30 ± 2.92**	Réunion	**96.80 ± 0.45**	92.42 ± 2.38
Argentina	**86.26 ± 0.47**	86.00 ± 4.40	Moldova	**90.18 ± 0.79**	78.11 ± 16.79
India	69.09 ± 1.74	**76.33 ± 7.77**	Indonesia	64.31 ± 0.84	**69.25 ± 2.81**
Romania	**94.49 ± 1.52**	90.52 ± 2.35	Bermuda	92.31 ± 0.60	**92.65 ± 1.84**
Slovakia	82.10 ± 0.98	**89.40 ± 5.09**	Malaysia	78.93 ± 3.78	**92.76 ± 2.55**
Hungary	**48.92 ± 3.59**	24.61 ± 3.35	South Africa	**95.31 ± 1.68**	92.75 ± 7.43
Japan	**41.41 ± 3.21**	33.34 ± 3.83	Latvia	**93.66 ± 0.64**	72.46 ± 5.77
Iceland	96.55 ± 1.20	**97.61 ± 0.98**	Kazakhstan	86.33 ± 3.06	**88.28 ± 11.32**
Venezuala	**94.87 ± 0.53**	89.82 ± 5.74	New Caledonia	**99.48 ± 0.15**	96.44 ± 5.64
Philippines	77.76 ± 3.97	**78.00 ± 11.75**	Estonia	**87.08 ± 1.89**	76.18 ± 1.62
Slovenia	95.37 ± 0.23	**96.47 ± 2.05**	Singapore	**86.42 ± 2.36**	83.23 ± 6.38
Ukraine	**92.99 ± 0.70**	90.86 ± 2.90	Bangladesh	78.61 ± 0.43	**79.77 ± 3.65**
Belarus	**91.08 ± 3.08**	90.16 ± 11.89	Paraguay	96.01 ± 1.23	**96.22 ± 1.78**
Serbia	**95.31 ± 0.48**	88.49 ± 7.05	Cyprus	**97.67 ± 0.34**	92.92 ± 6.94
Croatia	**94.59 ± 2.21**	88.17 ± 4.58	Bosnia	**84.04 ± 1.47**	80.53 ± 6.56
Greece	**81.98 ± 0.60**	35.30 ± 13.51	Ireland	**87.44 ± 0.69**	84.93 ± 2.85
New Zealand	94.27 ± 1.50	**97.77 ± 3.23**	Algeria	**85.37 ± 2.05**	79.66 ± 11.68
Portugal	**93.65 ± 0.46**	90.13 ± 4.47	Colombia	**87.81 ± 0.92**	87.60 ± 3.61
Bulgaria	**91.03 ± 2.07**	87.44 ± 11.94	Uzbekistan	**86.76 ± 1.13**	73.75 ± 3.42
Lithuania	**87.67 ± 3.05**	75.67 ± 2.19			

Agenda

Future Work

- Attention mechanism [?]
- Domain adaptation techniques (e. g. DANN) (CITE)

Conclusion

- Tackled the multinational address parsing problem with SOTA results.
- Shown that subword embeddings thelp to solve the multilingual aspect of our task.
- We explored the possibility of zero-shot transfer across countries and achieved interesting, but not yet optimal results.