Leveraging Subword Embeddings for Multinational Address Parsing

Marouane Yassine, David Beauchemin, François Laviolette, Luc Lamontagne

Department of Computer Science and Software Engineering, Université Laval

Introduction

We propose an approach in which we employ subword embeddings and a Recurrent Neural Network (RNN) architecture to build a single model capable of learning to parse addresses from multiple countries at the same time while taking into account the difference in languages and address formatting systems.

Motivations :

- Address parsing is an essential part of many applications, such as geocoding and record linkage.
- The only multinational solution takes a heavy pre-preprocessing and post-preprocessing step and a considerable amount of meta-data.

Related work :

- Sharma et al. (2018) : Parse monolingual address using a feedforward neural network.
- Mokhtari et al. (2019) : Parse monolingual address using different RNN architectures.

Goals :

- Provide multinational address parsing using a single model.
- Provide a more meaningful way to handle OOV words using subword embeddings.
- Evaluate the degree to which a model trained on countries' addresses data can perform well at parsing addresses from other countries (zero shot evaluation).

Subword Embeddings

Word embedding : vector representation of a word

- Non-contextual embeddings (e.g : Word2Vec, Glove)
- Contextual embeddings (e.g : ELMo, BERT)

Subword embedding : vector representation of a unit

- Character level
- Character n-grams (e.g : the bi-gram of "H1A 1B1" is H1, 1A, A1, 1B, B1) (e.g : fastText)
- Byte pair embeddings (e.g : the byte pair of "H1A 1B1" is {_H, 0, a, 0, b, 0}) (e.g : *BPEmb*)

Architecture

Embedding model : We compare two pre-trained embedding

- A fixed pre-trained monolingual fastText model (pre-trained on the French language) (**fastText**).
- A encoding of words using MultiBPEmb and merge the obtained embeddings for each word into one word embedding using a Bidirectional LSTM (Bi-LSTM) (hidden state dimension of 300) (**BPEmb**).

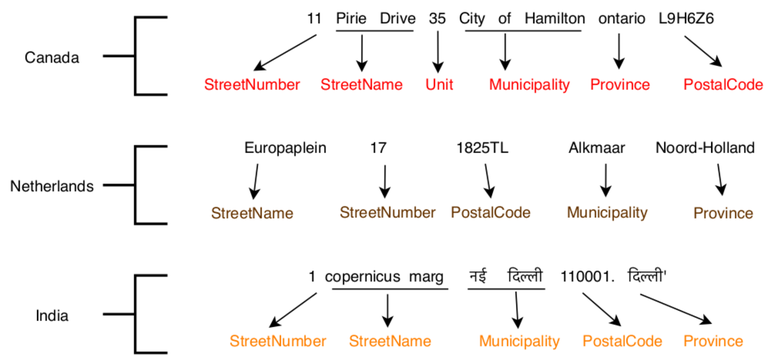
Tagging model : We use a Seq2Seq model consisting of

- a one-layer unidirectional LSTM encoder
- a one-layer unidirectional LSTM decoder
- a fully-connected linear layer to map the representation into the tag space dimensionality (i.e. the number of tags)
- a softmax activation.

Data

- Built using the open-source data of Libpostal.
- Contain 61 countries.
- We used eight tags : StreetNumber, StreetName, Unit, Municipality, Province, PostalCode, Orientation, and GeneralDelivery.

We use five different address patterns (one for each color) and some countries use more than one of these patterns (no color). The following figure presents three of them.



Experiments

Training data :

- 20 countries are used for multinational training with a sample size of 100,000 per country. The rest is used as a holdout.
- 41 countries are used for zero-shot transfer evaluation (never seen in training).

Training procedure :

- We trained each of our two models five times[1] (**fastText** and **BPEmb**).
- They were trained during 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.
- Trained using Poutyne.

Multinational Evaluation

Country	FastText	BPEmb	Country	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

- **FastText** gives the best performance across the board without considering the standard deviation.
- South Korean results are excellent despite the completely different alphabet.
- When using standard deviation, the **BPEmb** model achieves better results than **fastText** in most cases.
- We find that South Korea is the only country where a perfect accuracy (100 %) was achieved using **BPEmb** (3 out of 5).
- Randomly reordering 6000 South Korean addresses as either the first (red) or the second (brown) address pattern (equally divided between the two), the mean accuracy drops to 28.04% (the mean accuracy is of 12.29 % using a random tags procedure).

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			

- 50 % (19 out of 41) near state-of-the-art performance (> 90 %) for **fastText**. Or 35 % for **BPEmb**.
- 80 % (34 out of 41) good performance (> 80 %) for **fastText**. Or 65 % for **BPEmb**.
- The lowest results (below 70%) occur for countries where the address pattern and the country official language were not seen in the training data such as India, Hungary, and Japan.

Zero-shot Evaluation Discussion

For Hungary and Japan, the poorest results are mostly due to the address structure (blue), which is the near inverse of the two most present ones (red and brown) (never seen structure and language).

Kazakhstan, which uses the same address pattern as Japan, achieves better results. The main difference being the presence of one of the official language (Kazakh and **Russian**) in the training dataset.

India achieves almost 20% better results than Hungary and Japan, even if Hindi does not occur in the training dataset. Due to the use of a nearly identical address pattern as the first one (red). The only difference being the inversion of the province and the postal code.

Conclusion

- Tackled the multinational address parsing problem with SOTA results.
- We explored the possibility of zero-shot transfer across countries.

Future Work :

- Attention mechanism
- Domain-adversarial training techniques (e.g. DANN or ADANN)