Text segmentation with character-level text embeddings

Grzegorz Chrupała

http://grzegorz.chrupala.me

Introduction

- Most representations of text used in NLP are based on words.
- Distributional word classes or word embeddings successfully generalize over word forms.
- Words are not always available as units, and sometimes not the most appropriate level of granularity.
- Posts on a QA forum with segments of programming language example code embedded within the text.
- Use activations of hidden layer of SRN as text embeddings.

Simple Recurrent Networks

- Train 400-hidden unit SRNs on character sequences from Stackoverflow posts.

Example data

- Java - Convert String to enum

Say I have an enum which is just

```java
public enum Blah {
    A, B, C, D
}
```

and I would like to find the enum value of a string of for example “A” which would be Blah.A. How would it be possible to do this? Is `Enum.ValueOf()` the method I need? If so, how would I use this?

- Block and inline code fragments marked up with HTML tags.

Setup

- Collect questions from Stackoverflow.com between February and June 2011.
- Code blocks are delimited via HTML markup.
- Converted markup into labeled character sequences.
- Baseline: Trained CRF to predict labeling on raw text.
- Add SRN features on top of the baseline.

Evaluation

Embeddings trained on

- SMALL: 10 million characters
- LARGE: 456 million characters

Improvement from the SRN features largely due to their expressive power.

Embeddings boost performance as much as quadrupling the amount of labeled training examples does.

Results

Conclusion

- Created datasets and models for labeling code blocks in raw text.
- Showed that character-level text embeddings are useful representations for text segmentation.
- In future: joint model which learns to predict characters and their labels simultaneously.

References