

Text segmentation with character-level text embeddings

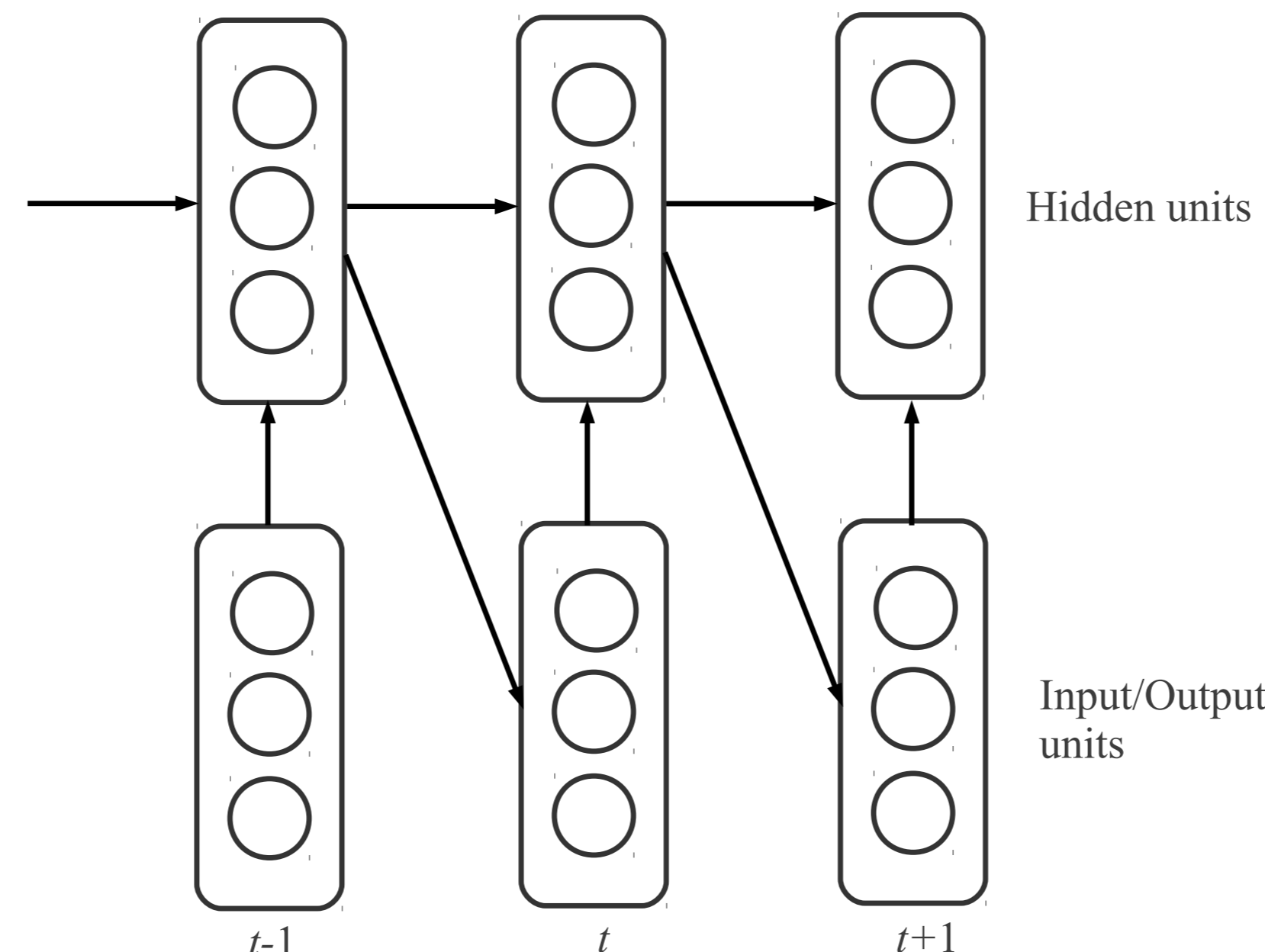
Grzegorz Chrupała

<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
public enum Blah {
 A, B, C, D
}
```
- 319 and I would like to find the enum value of a string  
60 of for example "A" which would be Blah.A. How  
would it be possible to do this?  
Is the Enum.valueOf() the method I need? If  
so, how would I use this?  
[java enums](#)
- ▶ 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.

### Baseline feature set

|          |   |  |    |    |     |
|----------|---|--|----|----|-----|
| Unigram  | t |  | p  | u  | b   |
| Bigram   |   |  | p  | pu |     |
| Trigram  |   |  |    | pu |     |
| Fourgram | t |  | pu |    | pub |
| Fivegram |   |  | t  |    | pub |

### Augmented feature set

For each of the  $K = 10$  most active units out of total  $J = 400$  hidden units, is the activation  $> 0.5$ ?

### Nearest neighbors in embedding space

```
n-laptop": {"last_share": 130738
ierre-pc": {"last_share": 130744
d-laptop": {"last_share": 130744
laptop": {"last_share": 13074434
erre-pc": {"last_share": 1307441
```

```
data table has integer values a
,2,3,4,5. For all these values I
ere i can add more connections s
eating lots of private methods a
or more different data sources c
```

```
e given URL.I'd like to change t
e = SqlPersist|||When I remove t
sources explaining how to save f
basic knowledge doesn't enable m
eDirectory, but I need to save t
```

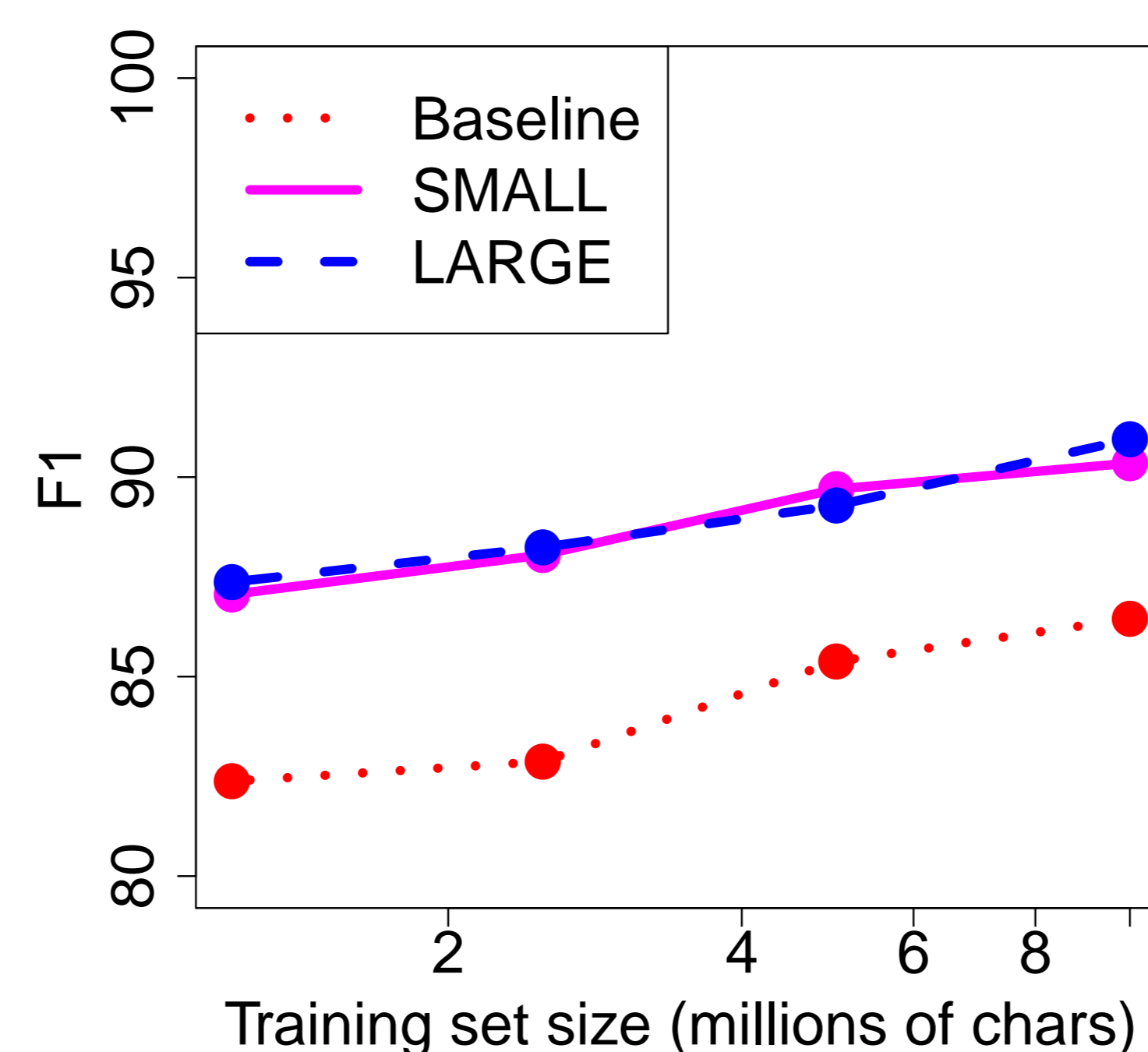
## Evaluation

- 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

- ▶ Elman, J. L. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- ▶ Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. Recurrent neural network based language model. In *Interspeech*, 2010.
- ▶ Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008.
- ▶ Sutskever, I., Martens, J., and Hinton, G. Generating text with recurrent neural networks. In *ICML*, 2011.