Deep learning for character-level text processing

Grzegorz Chrupała

Tilburg University

ATILA 2013

Small steps towards fully autonomous learning

Until recently **feature engineering** was a major bottleneck.

Until recently **feature engineering** was a major bottleneck.

Manually design, select and tune high-level, informative features.

Until recently **feature engineering** was a major bottleneck.

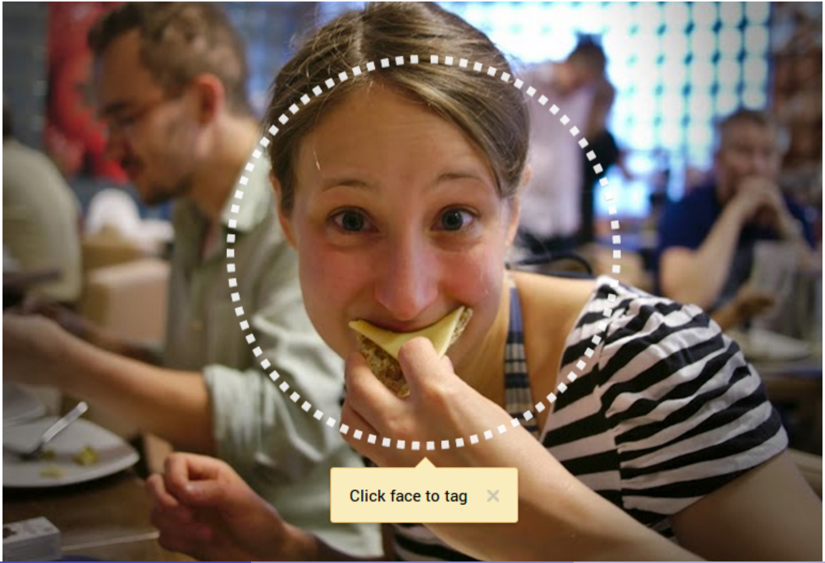
Manually design, select and tune high-level, informative features.

Feed these together with labels to shallow learners

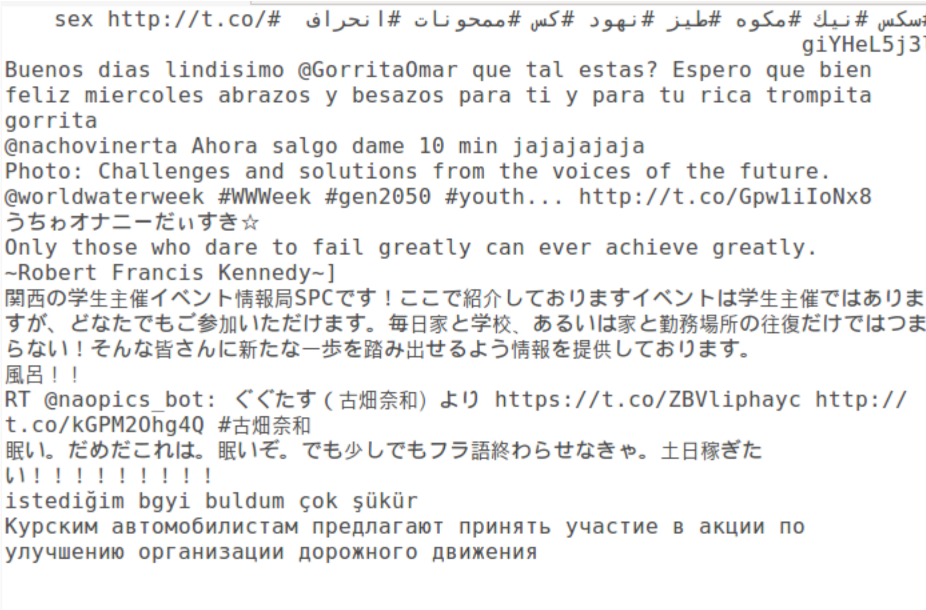
Recent focus on automatic learning of features

Neural networks in vision and speech

Machine vision: from pixels to labels via learned features

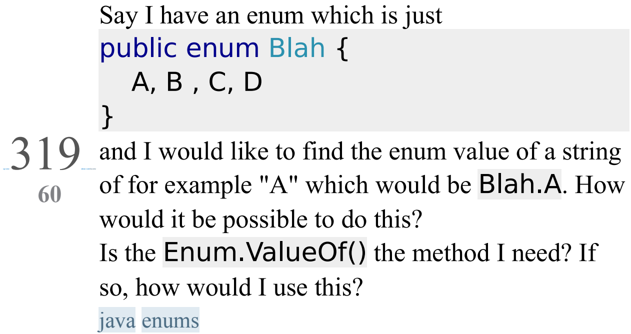


NLP: from characters to labels via learned features



#سكس #نيك #مكوه #طيز #نهود #كس #ممحونات #انحراف #sex http://t.co/giYHeL5j3
Buenos dias lindisimo @GorritaOmar que tal estas? Espero que bien
feliz miercoles abrazos y besazos para ti y para tu rica trompita
gorrita
@nachovinerta Ahora salgo dame 10 min jajajajaja
Photo: Challenges and solutions from the voices of the future.
@worldwaterweek #WWWeek #gen2050 #youth... http://t.co/Gpw1iIoNx8
うちゎオナニーだぃすき☆
Only those who dare to fail greatly can ever achieve greatly.
~Robert Francis Kennedy~]
関西の学生主催イベント情報局SPCです！ここで紹介しておりますイベントは学生主催ではありま
すが、どなたでもご参加いただけます。毎日家と学校、あるいは家と勤務場所の往復だけではつま
らない！そんな皆さんに新たな一歩を踏み出せるよう情報を提供しております。
風呂！！
RT @naopics_bot: ぐぐたす（古畑奈和）より https://t.co/ZBVliphayc http://
t.co/kGPM2Ohg4Q #古畑奈和
眠い。だめだこれは。眠いぞ。でも少しでもフラ語終わらせなきゃ。土日稼ぎた
い！！！！！！！！！！
istediğim bgyi buldum çok şükür
Курским автомобилистам предлагают принять участие в акции по
улучшению организации дорожного движения

Java - Convert String to enum



319

60

Say I have an enum which is just

```
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 the `Enum.ValueOf()` the method I need? If so, how would I use this?

java enums

Common approaches to representation learning for language

Common approaches to representation learning for language

- Word classes

Common approaches to representation learning for language

- Word classes
- Topics

Common approaches to representation learning for language

- Word classes
- Topics
- Word embeddings

Alternative

Alternative

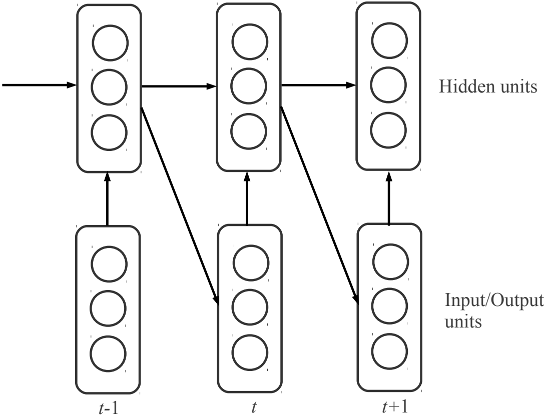
Learn text representations from raw character streams.

- Train a simple recurrent network to predict the next symbol in a sequence of characters (or bytes).

- Train a simple recurrent network to predict the next symbol in a sequence of characters (or bytes).
- Run trained network on new text.

- Train a simple recurrent network to predict the next symbol in a sequence of characters (or bytes).
- Run trained network on new text.
- Text representation: Activation of hidden layer at each position.

Finding Structure in Time, Elman 1990



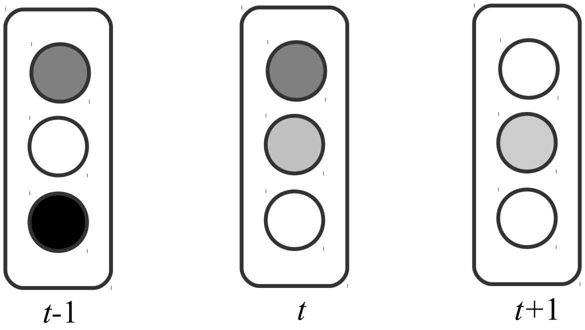
Hidden units

Hidden units

- Encode history

Hidden units

- Encode history
- Hopefully, generalize


t-1
t
t+1

- Train SRN on unlabeled text

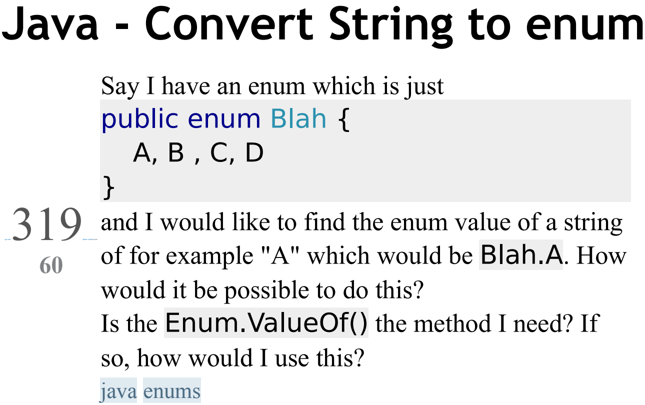
- Train SRN on unlabeled text
- Learn labels:

- Train SRN on unlabeled text
- Learn labels:
  - Extract simple **baseline** features from examples

- Train SRN on unlabeled text
- Learn labels:
    - Extract simple **baseline** features from examples
    - Run trained SRN on examples to get **learned** representations

- Train SRN on unlabeled text
- Learn labels:
  - Extract simple **baseline** features from examples
  - Run trained SRN on examples to get **learned** representations
  - Use a supervised shallow learner (e.g. CRF) with the union of **baseline** features and **learned** features

Segmenting STACKOVERFLOW posts



Java - Convert String to enum

319

60

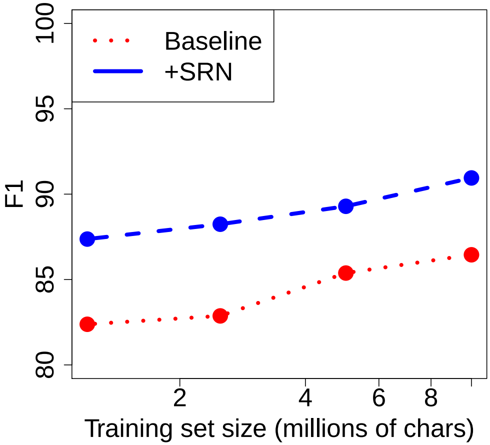
Say I have an enum which is just

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

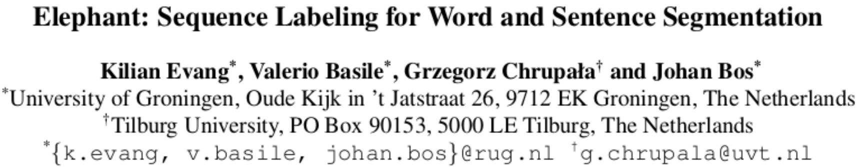
and I would like to find the enum value of a string of for example "A" which would be `Blah.A`. How would it be possible to do this?

Is the `Enum.ValueOf()` the method I need? If so, how would I use this?

java enums



Baseline
+SRN
F1
80
85
90
95
100
2
4
6
8
Training set size (millions of chars)



Elephant: Sequence Labeling for Word and Sentence Segmentation

Kilian Evang*, Valerio Basile*, Grzegorz Chrupała† and Johan Bos*
*University of Groningen, Oude Kijk in ’t Jatstraat 26, 9712 EK Groningen, The Netherlands
†Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
*{k.evang, v.basile, johan.bos}@rug.nl †g.chrupala@uvt.nl

```
It didn’t matter if the faces were male,
SIOTIITIIOTIIIIIOTIOTIIOTIIIIOTIIIOTIIITO
female or those of children. Eighty-
TIIIIIOTIOTIIIIOTIOTIIIIIIIITOSIIIIIIO
three percent of people in the 30-to-34
IIIIIOTIIIIIIOTIOTIIIIIOTIOTIIOTIIIIIIIO
year old age range gave correct responses.
TIIIOTIIOTIIOTIIIIOTIIIOTIIIIIIOTIIIIIIIIIT
```

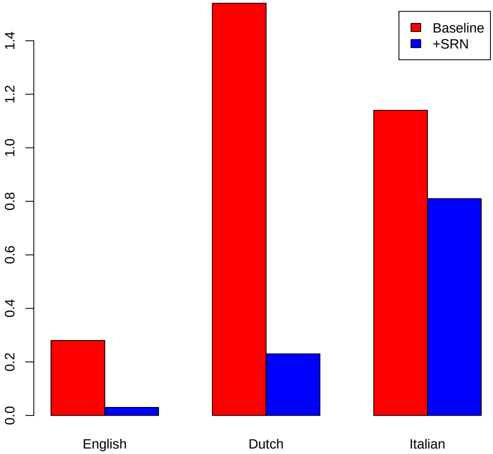
- Character-wise sequence labeler (CRF)

- Character-wise sequence labeler (CRF)
- Baseline features: characters and their Unicode categories in a window around current position

- Character-wise sequence labeler (CRF)
- Baseline features: characters and their Unicode categories in a window around current position
- SRN features: discretized activations of hidden layer

- Character-wise sequence labeler (CRF)
- Baseline features: characters and their Unicode categories in a window around current position
- SRN features: discretized activations of hidden layer
- English, Dutch and Italian

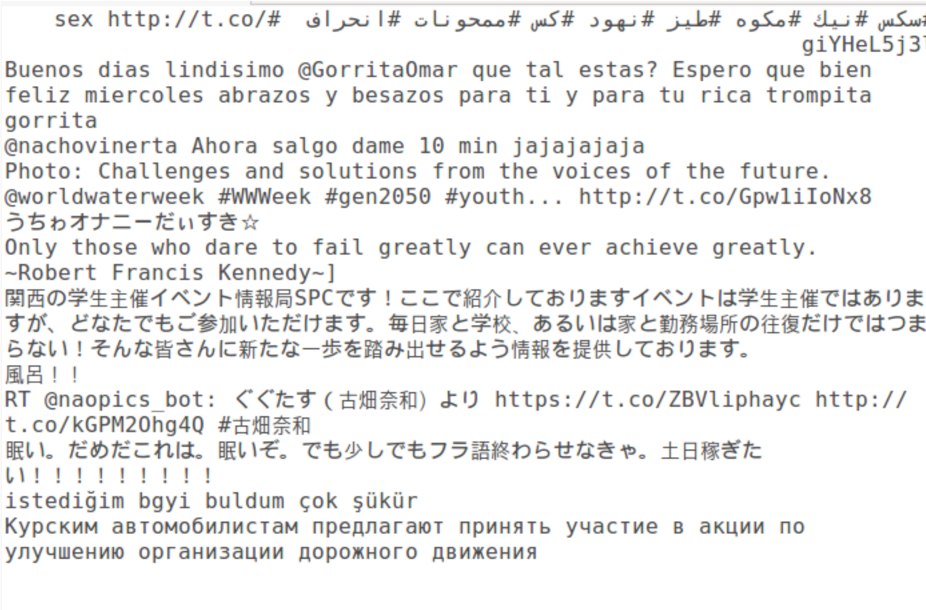
Errors per thousand characters



	Ms. Hughes will joi
	SIIOSIIIIIOTIIIOTII
+SRN	SIIOTIIIIIOTIIIOTII
	$ 3.9 trillion by t
	TOTTIOTIIIIIIIOTIOT
+SRN	TOTIIOTIIIIIIIOTIOT
	prof. Teulings het
	TIIITOSIIIIIIIOTIIO
+SRN	TIIIIOTIIIIIIIOTIIO
	bleek 0,4 procent
	OTIIIIOTTIOTIIIIIIO
+SRN	OTIIIIOTIIOTIIIIIIO
	per costringerlo al
	TIIOTIIIIIIIIIIIOTI
+SRN	TIIOTIIIIIIIIITIOTI

Work in progress: labeling tweets

Add / suggest hashtags to tweets or other user generated content



```
#سكس #نيك #مكوه #طيز #نهود #كس #ممحونات #انحراف #sex http://t.co/
giYHeL5j3
Buenos dias lindisimo @GorritaOmar que tal estas? Espero que bien
feliz miercoles abrazos y besazos para ti y para tu rica trompita
gorrita
@nachovinerta Ahora salgo dame 10 min jajajajaja
Photo: Challenges and solutions from the voices of the future.
@worldwaterweek #WWWeek #gen2050 #youth... http://t.co/Gpw1iIoNx8
うちゎオナニーだぃすき☆
Only those who dare to fail greatly can ever achieve greatly.
~Robert Francis Kennedy~]
関西の学生主催イベント情報局SPCです！ここで紹介しておりますイベントは学生主催ではありま
すが、どなたでもご参加いただけます。毎日家と学校、あるいは家と勤務場所の往復だけではつま
らない！そんな皆さんに新たな一歩を踏み出せるよう情報を提供しております。
風呂！！
RT @naopics_bot: ぐぐたす（古畑奈和）より https://t.co/ZBVliphayc http://
t.co/kGPM20hg4Q #古畑奈和
眠い。だめだこれは。眠いぞ。でも少しでもフラ語終わらせなきゃ。土日稼ぎた
い！！！！！！！！！！
istediğim bgyi buldum çok şükür
Курским автомобилистам предлагают принять участие в акции по
улучшению организации дорожного движения
```

Tweets notoriously hard to deal with using traditional NLP pipelines

Tweets notoriously hard to deal with using traditional NLP pipelines

- Multitude of languages and scripts

Tweets notoriously hard to deal with using traditional NLP pipelines

- Multitude of languages and scripts
- Nonstandard spelling and punctuation

Tweets notoriously hard to deal with using traditional NLP pipelines

- Multitude of languages and scripts
- Nonstandard spelling and punctuation
- Abbreviations, slang etc.

Can features learned from raw bytes strings help?

Train SRN on a stream of tweets.

Train SRN on a stream of tweets.

- Use common hashtags to learn to label

Train SRN on a stream of tweets.

- Use common hashtags to learn to label
- Single-pass online learner

Train SRN on a stream of tweets.

- Use common hashtags to learn to label
- Single-pass online learner
- Baseline character n-gram features + SRN features

Train SRN on a stream of tweets.

- Use common hashtags to learn to label
- Single-pass online learner
- Baseline character n-gram features + SRN features
- Activation vectors recorded at valleys in entropy profile, and averaged together

Very preliminary results

on 5000 items

Features	Mean Average Precision
char. n-grams	31.1
char. n-grams +SRN	**34.8**

Thank you

Sample of nearest neighbors according to cosine of the hidden layer activation in a span of 10.000 characters

```
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
```

Generated random text

```
I·only·make·event·glds.

so,·on·the·cell·proceedclicks·like·completed,·with·color?

····st·potention,
‘column’]HeaderException=ID·=·new·Put="True"·MetadataTemplate,
·grwTrowerRow="SELECTEMBRow"·on?

All·clearBeanLockCollection="#7293df3335b-E9"·/&gt;
············&lt;Image:DataKey="BackgroundCollectionC2UTID"·
onclick="Nore"·
```

Labels

Block

w	r	o	n	g	?	¶	t	r	y
O	O	O	O	O	O	O	B-BL	I-BL	I-BL

Inline

e	r	·	.	.	/	i	m	g
O	O	O	B-IN	I-IN	I-IN	I-IN	I-IN	I-IN

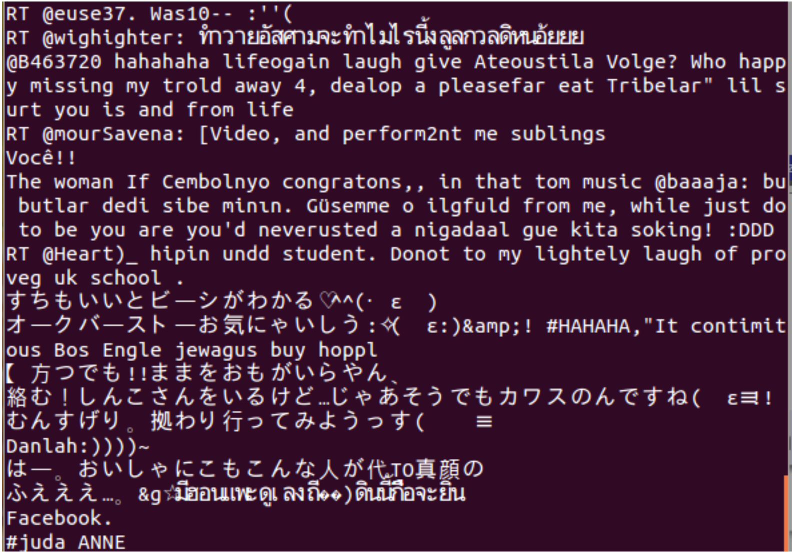
Baseline feature set

...wrong?¶try {...

Unigram	n g ? ¶ t
Bigram	g? ?¶
Trigram	g?¶
Fourgram	ng?¶ g?¶t
Fivegram	ng?¶t

Augmented feature set

- Baseline features
- 400-unit hidden layer activation
    - For each of 10 most active units
        - Is the activation > 0.5?



```
RT @euse37. Was10-- :''(
RT @wighighter: ทำวายอัสศามจะทำไม่ไรนี้งลูลกวลดิหน้อ้ยยย
@B463720 hahahaha lifeogain laugh give Ateoustila Volge? Who happ
y missing my trold away 4, dealop a pleasefar eat Tribelar" lil s
urt you is and from life
RT @mourSavena: [Video, and perform2nt me sublings
Você!!
The woman If Cembolnyo congratons,, in that tom music @baaaja: bu
 butlar dedi sibe minın. Güsemme o ilgfuld from me, while just do
 to be you are you'd neverusted a nigadaal gue kita soking! :DDD
RT @Heart)_ hipin undd student. Donot to my lightely laugh of pro
veg uk school .
すちもいいとビーシがわかる♡^^(· ε  )
オークバースト—お気にゃいしう:✧(  ε:)&amp;! #HAHAHA,"It contimit
ous Bos Engle jewagus buy hoppl
【 方つでも!!ままをおもがいらやん、
絡む！しんこさんをいるけど…じゃあそうでもカワスのんですね(  ε≡!
むんすげり。 拠わり行ってみようっす(    ≡
Danlah:))))~
はー。おいしゃにこもこんな人が代TO真顔の
ふええええ…。&g☆มีฮอนแพะดูเ ลงถี◆◆)ดินนี้ก็อจะยิน
Facebook.
#juda ANNE
```