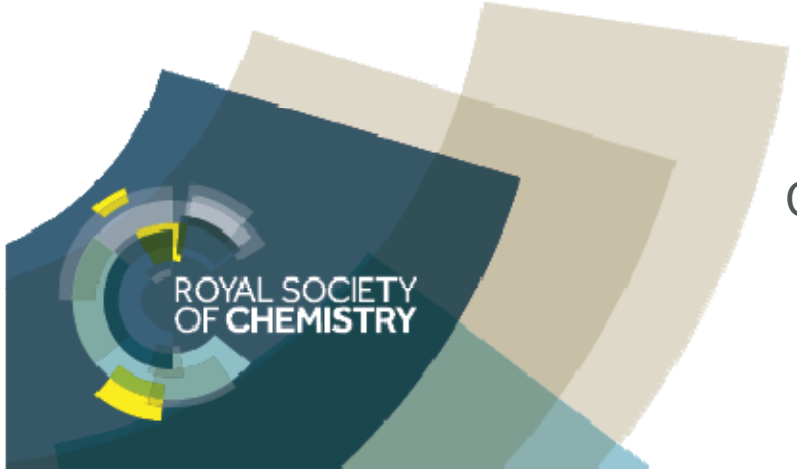
Deep learning and chemical data



Colin Batchelor, Nicholas Bailey, Peter Corbett, Aileen Day, Jeff White and John Boyle

2018-06-15

Overview

- Who we are
- Chemical structure elucidation from NMR spectra
- Chemical named entity recognition and recurrent neural networks
- Relation extraction and transfer learning

Data Science at the Royal Society of Chemistry

- Helping other teams make evidence-based decisions.
- Making RSC increasingly data-informed.
- Developing new ways of handling chemical science information.

Data Science at the Royal Society of Chemistry

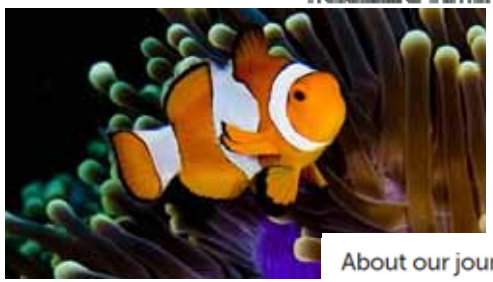
- Helping other teams make evidence-based decisions.
- Making RSC increasingly data-informed.
- Developing new ways of handling chemical science information.

Chemical data at the RSC

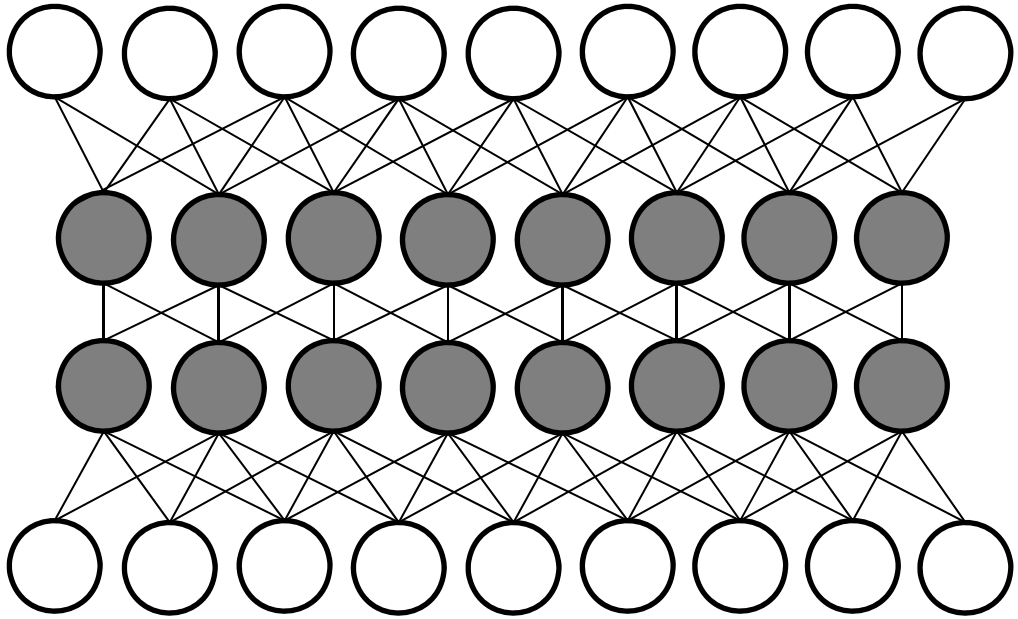
- Chemical structures
- Spectra
- Reaction schemes
- Unstructured text

Chemical data at the RSC

- Chemical structures
- Spectra
- Reaction schemes
- Unstructured text



Deep learning



Input vector

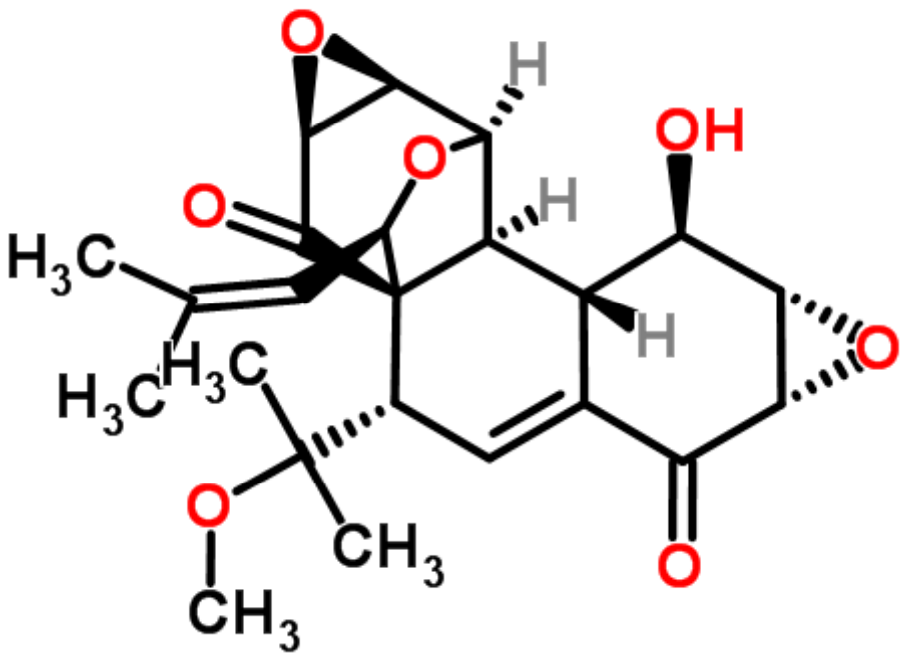
Hidden layers

“Feature engineering”
happens here through training.

Output vector

Part I
NMR (Summer 2017)


http://mushroomobserver.org/observer/show_user/2250


O
H
OH
O
O
H
H
H3C
H3C
H3C
O
CH3
CH3
O
O

Task

Take a ^{13}C-NMR spectrum and return the functional groups in the molecule.

How do we do this?

Source data

Real: 34629 spectra and structures | nmrshiftdb2withsignals.sd (NMRShiftDB2, http://nmrshiftdb.nmr.uni-koeln.de)

Synthetic: 24900 spectra and structures | MarinLit Database (Royal Society of Chemistry, http://pubs.rsc.org/marinlit)

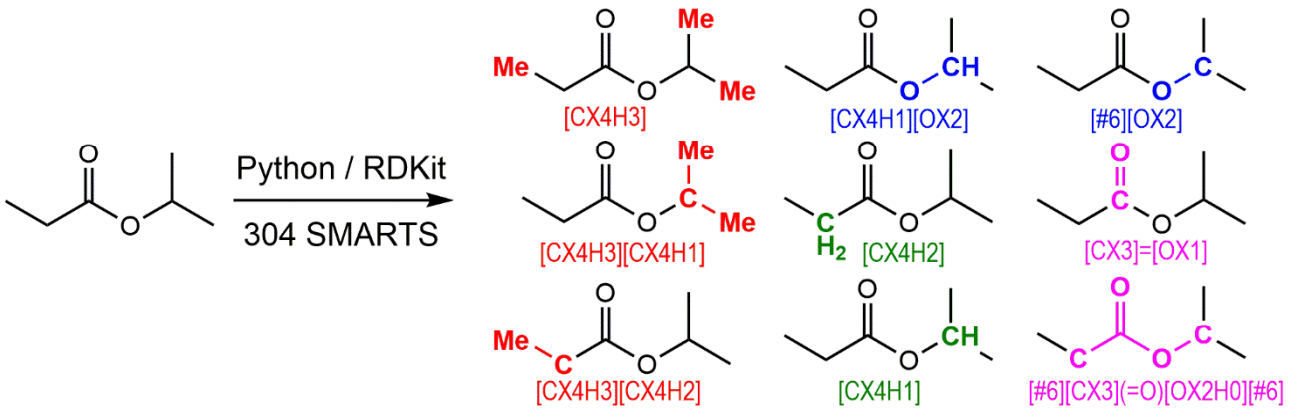
Input spectra

Shift range: $310 \geq \delta > -40$

Resolution: 0.1 ppm

Vectors $[\, x_0\ x_1\ \ldots\ x_{3499}\,]$

Output



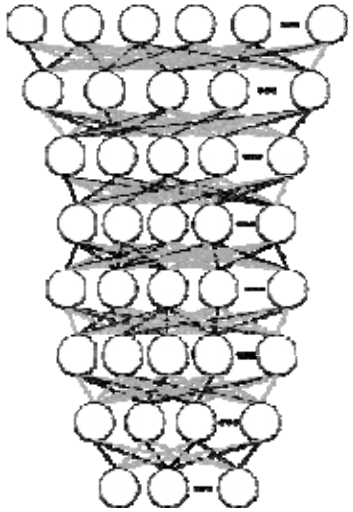
0 1 2 3 4 5 6 7 8 9 10 11 18 19 20 21 26 27 28 29 68 69

Number of groups → [3, 0, 0, 0, 0, 2, 1, 0, 0, 1, 1, 0, … 0, 1, 1, 0, … 0, 1, 1, 0, … 0, 0]

Architectures

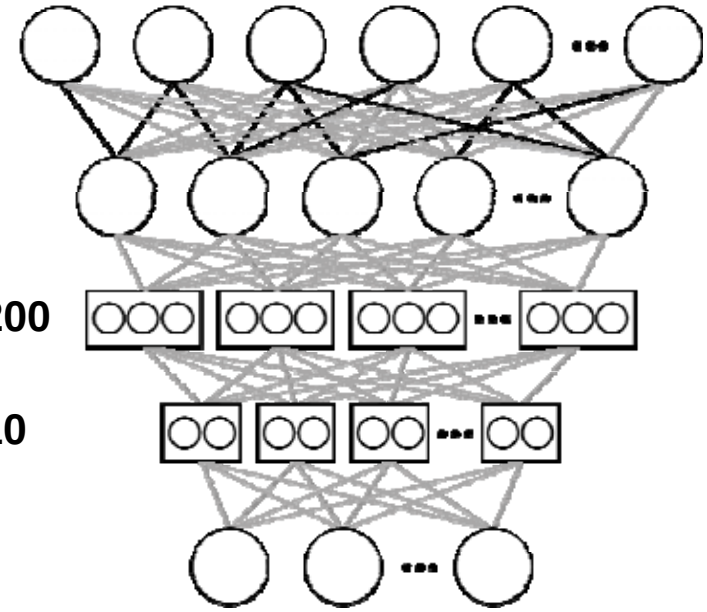
Dense

Input | 3500
3000
2000
1000
2000
1000
500
Output | 70



Convolution

Input | 3500
3000
512@200
512@10
Output | 70

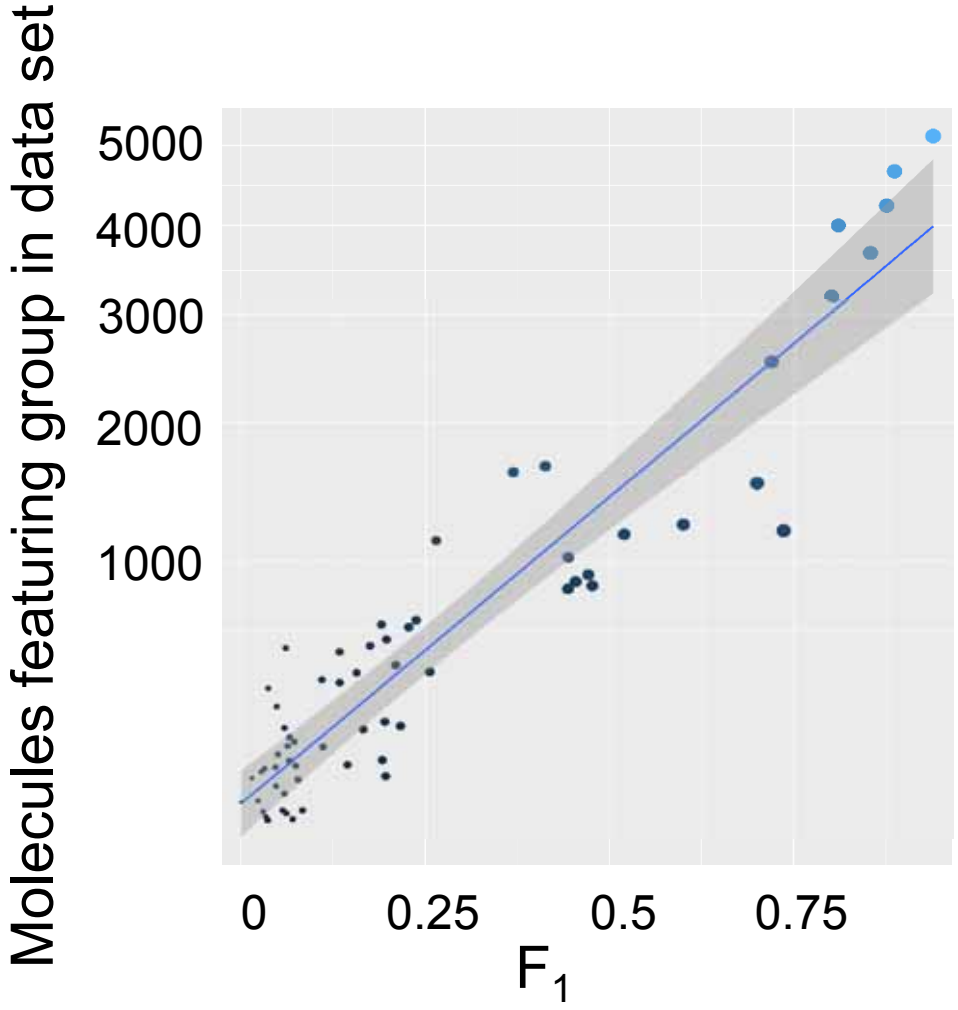


`import`

- RDKit for cheminformatics
- Keras for deep learning
- Theano

Results based on real spectra

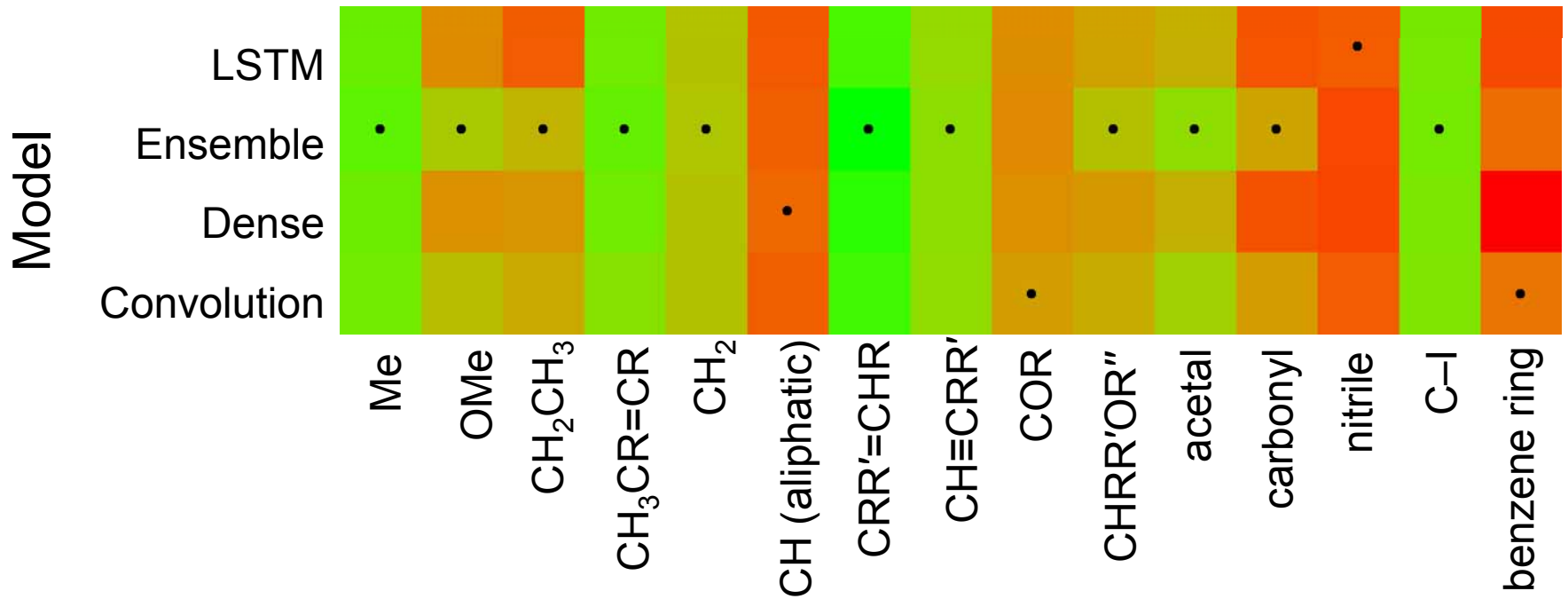
There is a clear relationship between the number of times molecules featuring a particular group appear in the training data and the F_1 score. More data is needed for better training for all groups.



Results (micro-F)

Model	Precision	Recall	F_1
Convolution	0.7391	0.7628	**0.7508**
Dense	0.7402	0.7340	0.7370
Ensemble	0.6973	**0.8049**	0.7472
LSTM	**0.7592**	0.6984	0.7275
Baseline: random forest	0.7233	0.5633	0.6334

Breakdown


F1
0.4
0.6
0.8
• = Best result
Model
LSTM
Ensemble
Dense
Convolution
Me
OMe
CH2CH3
CH3CR=CR
CH2
CH (aliphatic)
CRR′=CHR
CH≡CRR′
COR
CHRR′OR″
acetal
carbonyl
nitrile
C–I
benzene ring

NMR conclusions

- Good performance for some functional groups.
- Natural product structures are extremely challenging.
- This could really benefit from more training data.

Part II
Chemical named-entity recognition (April 2017)

BioCreative V.5

Chemical Entity Mentions in Patents

- A61K 31 - Medicinal preparations containing organic active ingredients
- A61P - Specific therapeutic activity of chemical compounds or medicinal preparations
- 21000 training abstracts, 9000 test abstracts

The BioCreative V.5 evaluation workshop: tasks, organization, sessions and topics. Krallinger et al. *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop, 8-10*
Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. Pérez-Pérez et al. *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop, 11-18*

Why is it hard?

- Ambiguity: “lead”, “K”, many acronyms
- Out-of-dictionary terms
- Multiword terms
- Tokenisation
- Vagueness of task
  - “calcium ion” vs “calcium ion”

Task

Input: unstructured text:

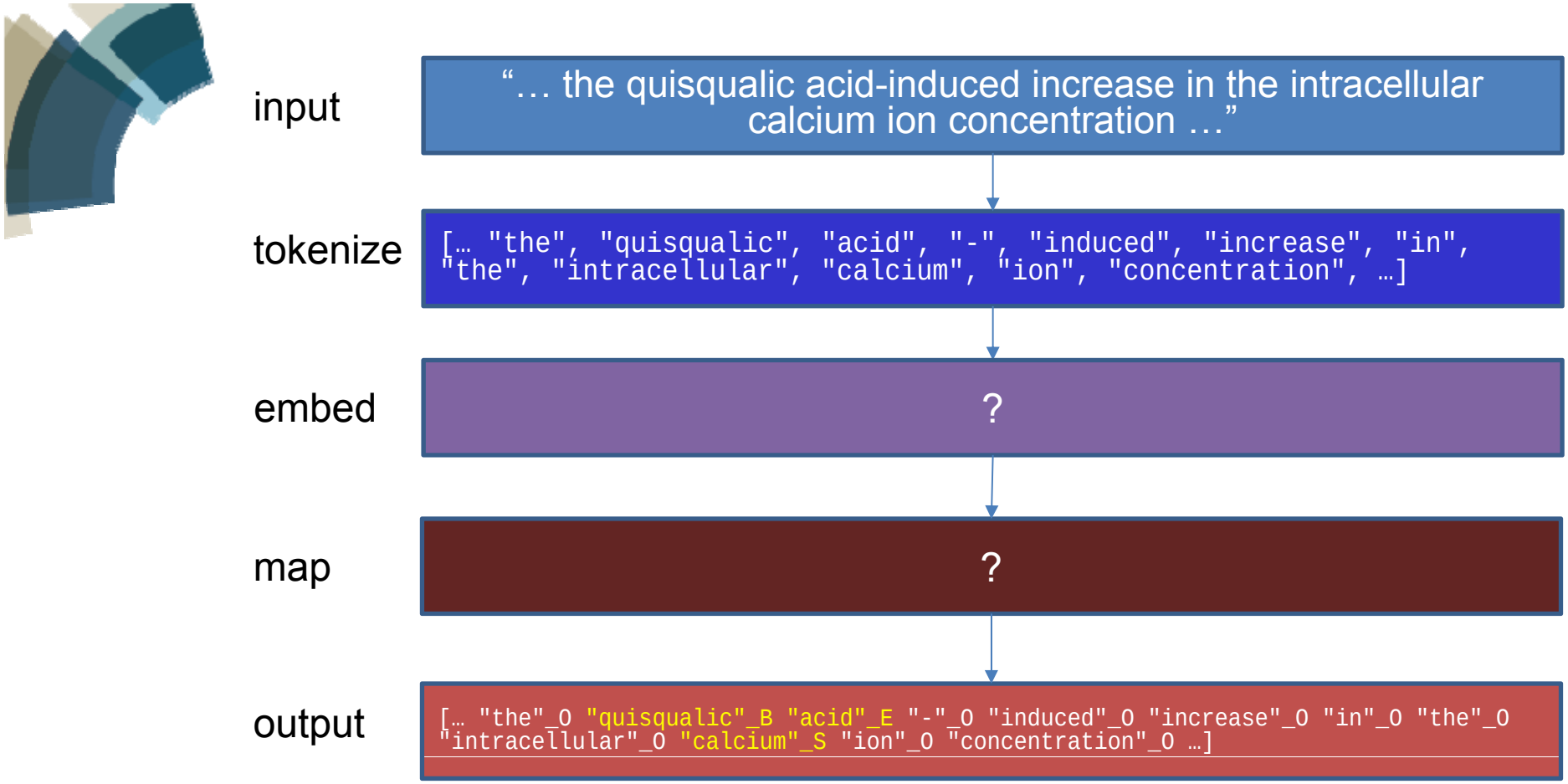
“… the quisqualic acid-induced increase in the intracellular calcium ion concentration …”

Output: character positions of beginnings and ends of chemical named entities.

Limits to performance

- Training data from manual corpus annotation
- Typical ceiling of 90-93% *F* for inter-annotator agreement
  - often ~70% *F* for "naïve annotators" with no guidelines

Peter Corbett, Colin Batchelor, and Simone Teufel. **Annotation of chemical named entities.** *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, 2007.



O: outside **B**: begin **I**: inside **E**: end **S**: singleton

Embeddings

- word2vec and GloVe are examples of word embeddings
- human-language specific
- high-dimensional (often 300) vector per token
- Similar to LSA *etc.*
- Trainable inside neural network

Two systems

"Traditional"	"Minimalist"
Token level – modified Oscar tokeniser	Character level – no tokeniser
Rich feature set + token embeddings	Character embeddings only
Uses external resources (e.g ChEBI)	No external resources
1 recurrent layer	3 recurrent layers
Trains in hours	Trains in days
Large model file	Small model file

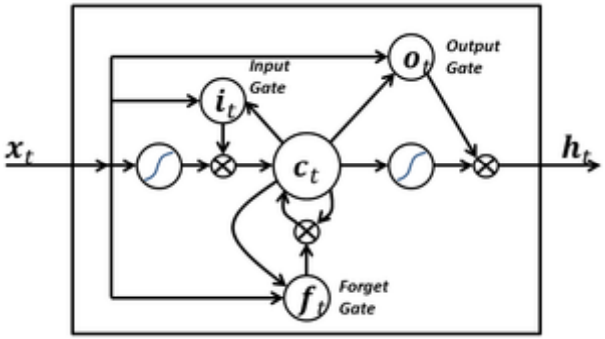
Features taken from:
Cascaded classifier for confidence-based chemical named entity recognition.
Peter Corbett and Ann Copestake, *BMC Bioinformatics,* 2008, **9(Suppl 11)**, S4.

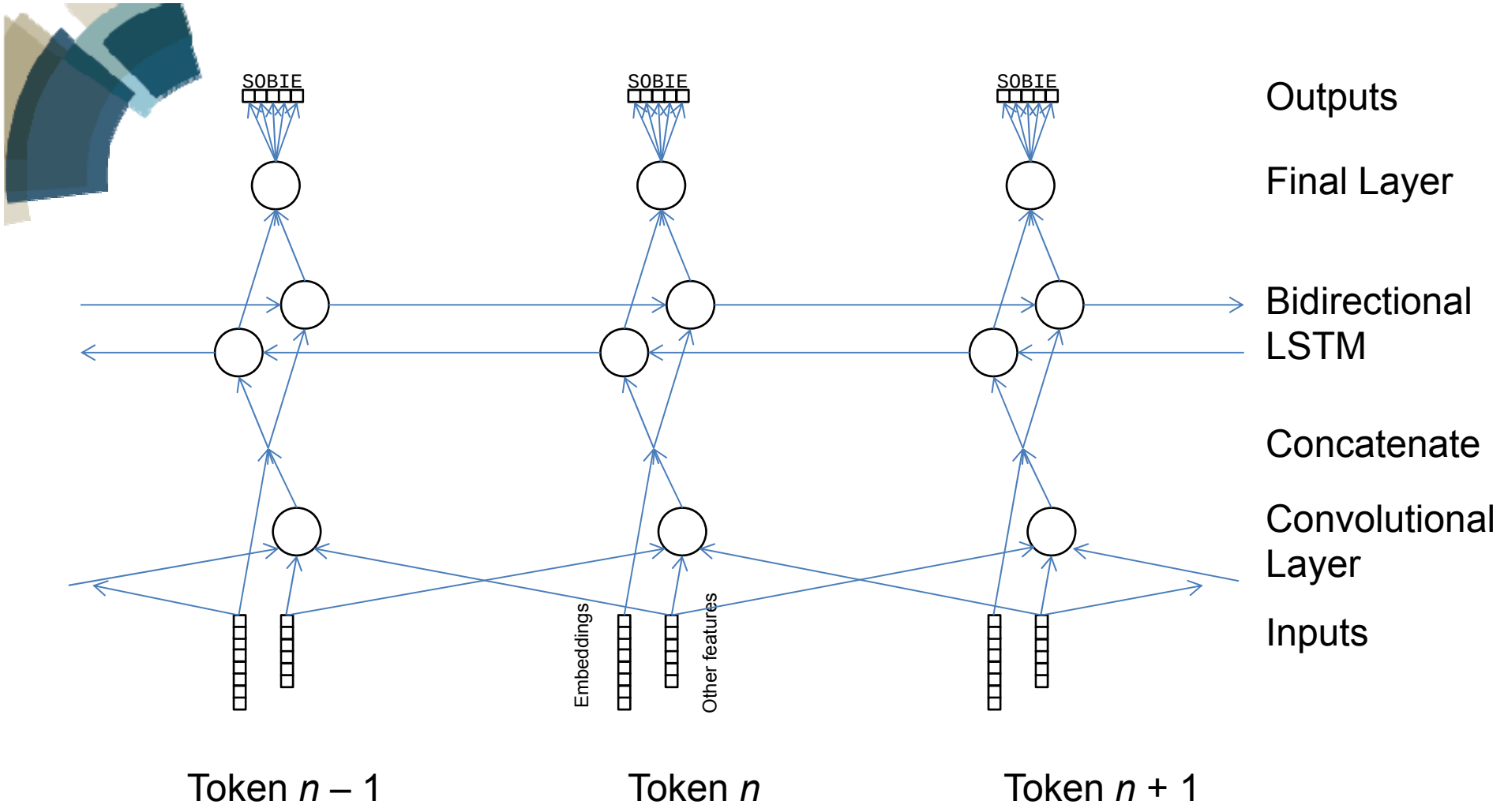
Feature examples

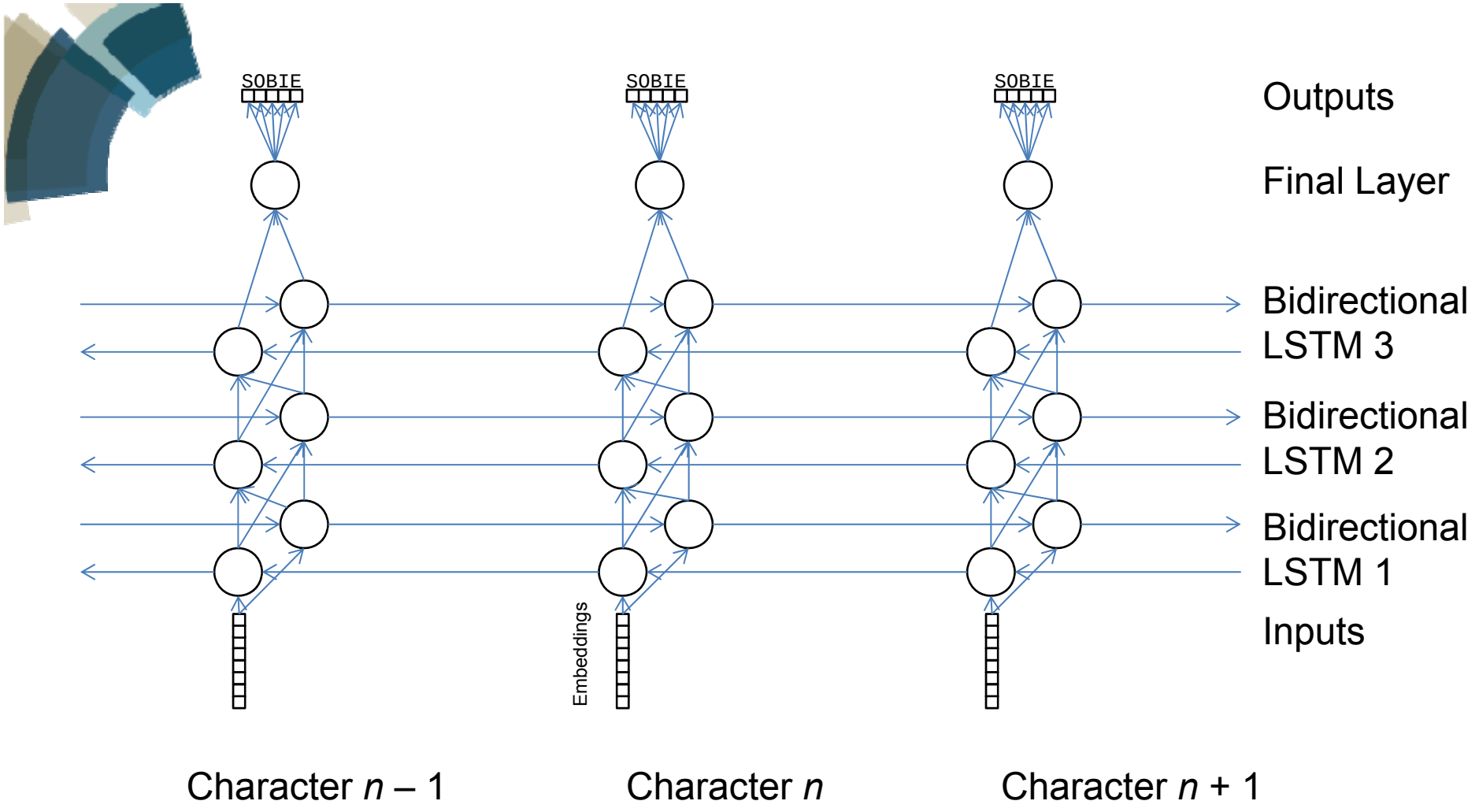
Feature	Examples
Regex	`/^.*[a-z][a-z].*$/` `/^.*([A-Za-z].*[0-9]\|[0-9].*[A-Za-z]).*$/`
In a list?	ChEBI, /usr/share/dict/words
character n-gram	thyl
Prefix or suffix?	`^tri` `one$`
Length	

Recurrent Neural Networks

- Neural network where outputs feed into inputs – directed cyclic graph
- Equivalent to multiple repeats of the network, as a DAG – very deep network
- LSTM – units have internal structure, has “forget gates” to avoid “vanishing gradient problem”




SOBIE
SOBIE
SOBIE
Outputs
Final Layer
Bidirectional LSTM
Concatenate
Convolutional Layer
Inputs
Embeddings
Other features
Token n – 1
Token n
Token n + 1


SOBIE
SOBIE
SOBIE
Outputs
Final Layer
Bidirectional LSTM 3
Bidirectional LSTM 2
Bidirectional LSTM 1
Inputs
Embeddings
Character n – 1
Character n
Character n + 1

`import`

- chemtok for tokenization
- Keras
- TensorFlow
- scikit-learn
- h5py

Results

System	Official F	Official P	Official R	Internal F	Internal P	Internal R
Traditional	0.8919	0.8867	0.8971	0.8703	0.8648	0.8758
Minimalist	0.8901	0.8865	0.8936	0.8664	0.8479	0.8858
Ensemble	**0.9032**	**0.9002**	**0.9062**	**0.8807**	**0.8646**	**0.8976**

Chemlistem – chemical named entity recognition using recurrent neural networks, P. Corbett and J. Boyle, *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, 61–68.

BioCreative V.5 top three

Team	Method	F_1
Buzhou Tang group (Shenzhen)	Bidirectional LSTM (word embeddings and character embeddings)	0.9031
Zhihao Yang group (Dalian)	Bidirectional LSTM plus CRF (word embeddings and character embeddings)	0.9042
RSC	Bidirectional LSTM (ensemble of word embeddings and character embeddings)	0.9032

Part III
Relationship extraction (November 2017)

Chemical–protein relations

- Key to medicinal chemistry
- Chemical/drug = small molecule (not a protein)
- Gene makes protein (“gene product”), often referred to by same name

- Corpus taken from PubMed and manually annotated.

http://www.biocreative.org/tasks/biocreative-vi/track-5/

What did we expect?

- No precisely comparable task
- BioCreative chemical–disease winning *F*-score 57.03%
- BioCreative protein–protein interaction winning *F*-score 55%.

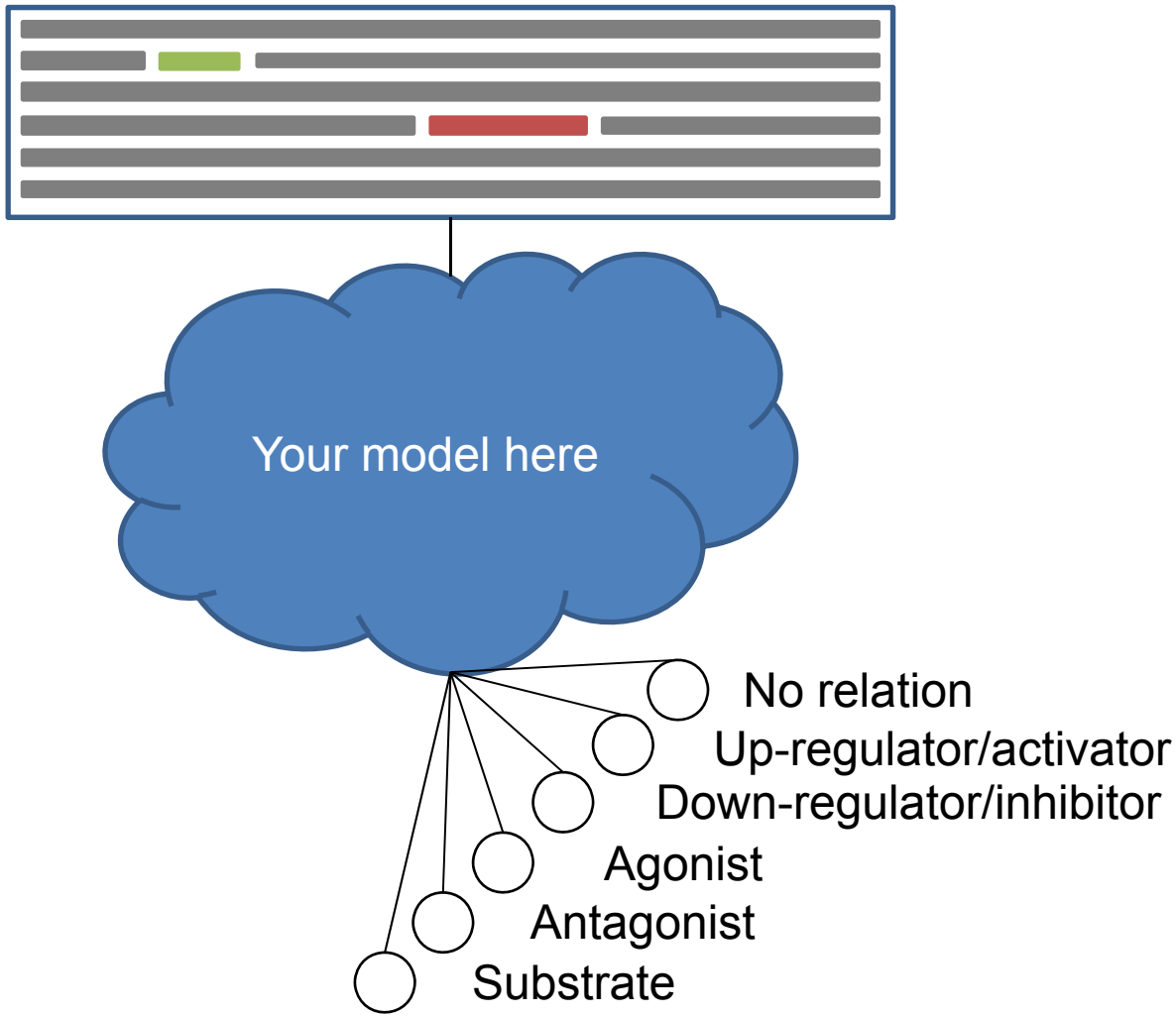
- Significantly lower than named entity recognition!

CHEMPROT example

```
15110853    CPR:4  Y     INHIBITOR    Arg1:T11    Arg2:T14
15110853    T11    CHEMICAL    759    772    methazolamide
15110853    T14    GENE-N 599    607    human CA

15110853    … Some of these derivatives showed good
inhibitory potency against two human CA isozymes involved in
important physiological processes, CA I, and CA II, of the
same order of magnitude as the clinically used drugs
acetazolamide and methazolamide.
```

Dataset	Size
Training	1020
Development	612
Test	3399



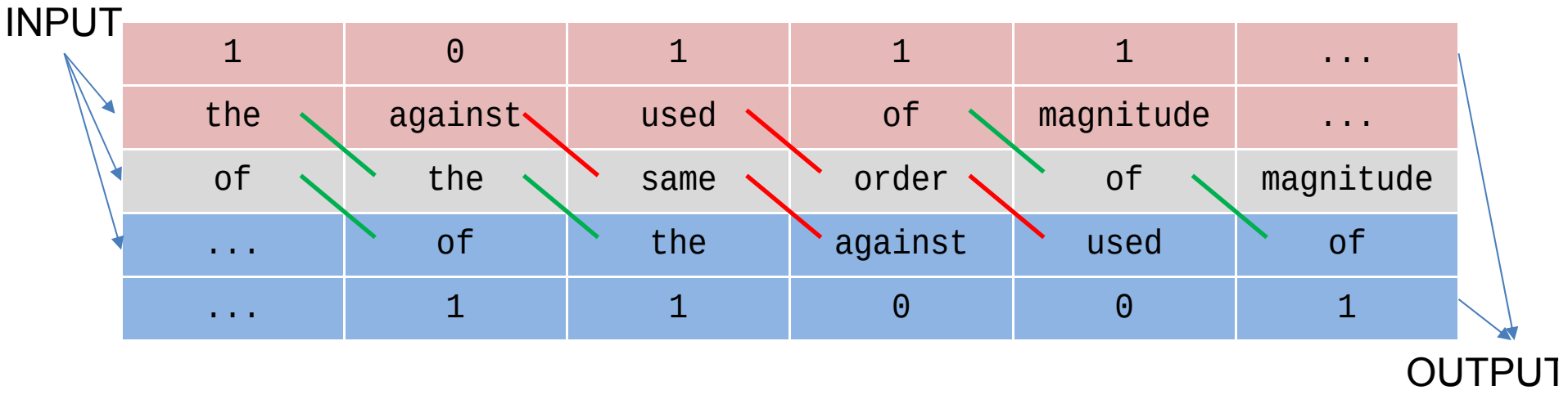
Transfer learning

Find a task a part of the network can do with unlabelled data, train it on that, then incorporate that part into the final network

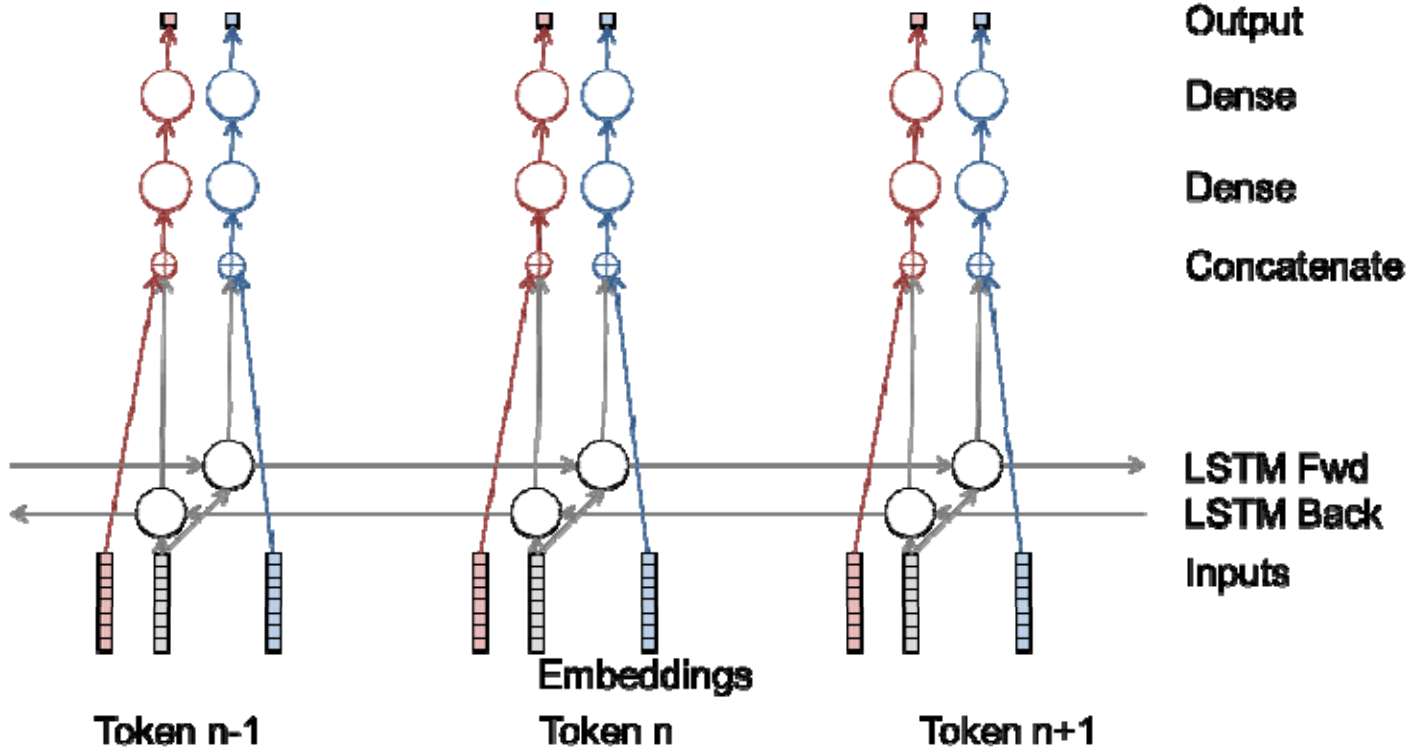
- Predict the next word (or previous word)
- “Is this word the next one”?
- PubMed abstracts, RSC papers, pharma patents to train GLoVE embeddings

Pretraining

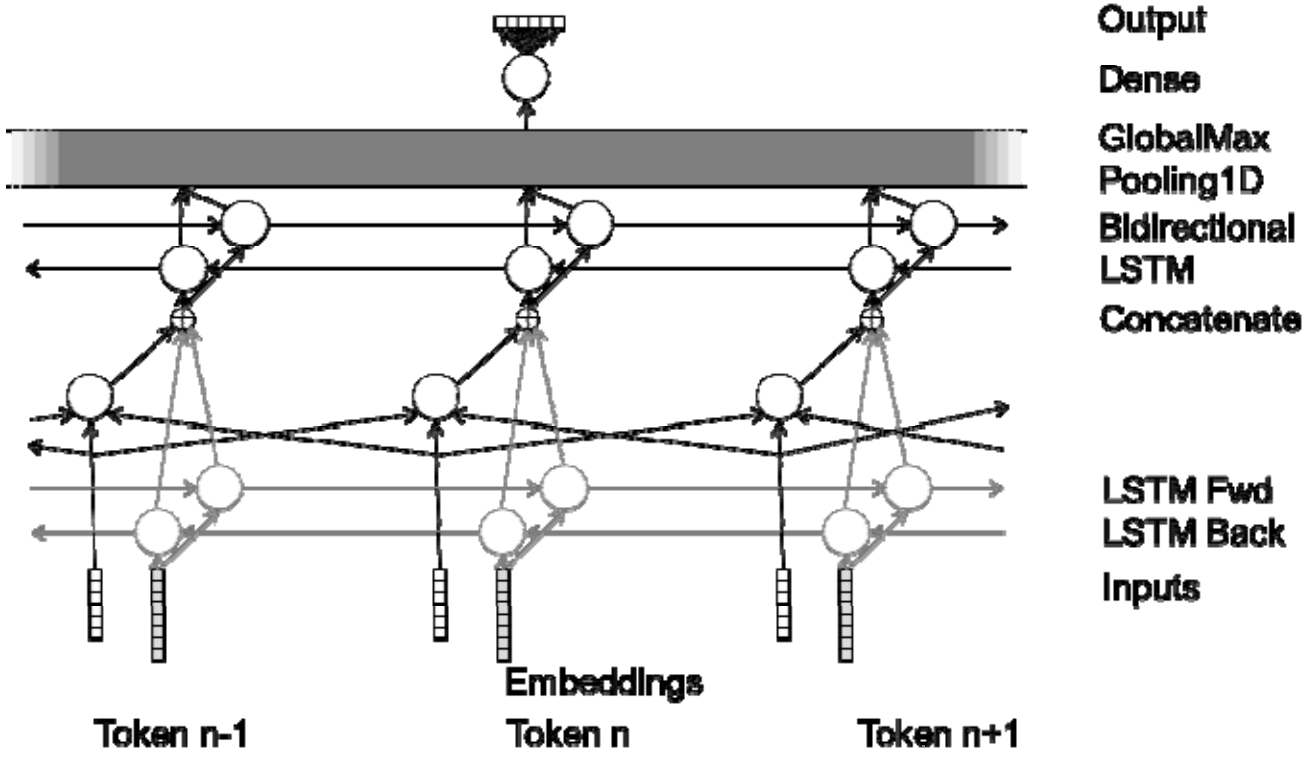
- Train system to predict next token given token and all previous, and previous token given token and all next.
- Generate sequences shifted 1 token left and right, half words of the replaced with random, and 1/0 to say whether word was replaced.



Pretraining network



Recognition network



`import`

- chemtok for tokenization
- Keras
- TensorFlow
- scikit-learn
- GloVe

Results

Corpus	Precision	Recall	*F*
Development	**0.5652**	**0.7042**	**0.6271**
Competition entry	0.5610	0.6784	0.6141

Ablation studies

Run	Precision	Recall	*F*
Random embeddings	0.4505	0.5066	0.4770
Public GloVe embeddings	0.6169	0.5696	0.5923
Chemically-trained GloVe embeddings	0.6297	0.5725	0.5997
Chemically-trained GloVe embeddings + transfer learning	**0.5652**	**0.7042**	**0.6271**

CHEMPROT overall results

Team	# runs	Best precision	Best recall	Best F
Peng (NCBI, NLM, NIH)	5	**0.7437**	0.5735	**0.6410**
Corbett (RSC)	1	0.5610	**0.6784**	0.6141
Mehryary (Turku)	3	0.6608	0.6006	0.6099
Lim (Korea University)	2	0.6760	0.5194	0.5853
Lung (Florida State)	2	0.6352	0.5121	0.5671

Peng *et al.*: **https://arxiv.org/abs/1802.01255**

Conclusions

- Character-based deep-learning models can perform at very close to human levels for chemical named-entity recognition.
- Relation extraction and NMR spectra are much harder tasks for deep learning.
- More data will mean better predictions.

References

https://bitbucket.org/rscapplications/chemlistem

```
pip install chemlistem
```


Chemlistem - chemical named entity recognition using recurrent neural networks. Corbett and Boyle 2017. *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop, 61-68.*

Full writeup: *J. Cheminf.*, submitted.

Improving the learning of chemical-protein interactions from literature using transfer learning and word embeddings. Corbett and Boyle 2017. *Proceedings of the BioCreative VI Workshop, 180-183.*

Full writeup: *Database,* in press.

Read the latest issue at
rsc.li/molecular-engineering

Sign up for issue alerts at
rsc.org/alerts

@RSC_MolEng

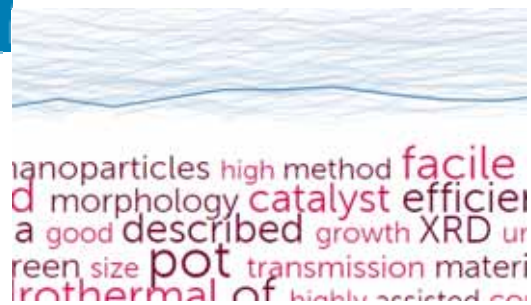
Machine learning special issue:
https://rsc.li/msde-machine-learning

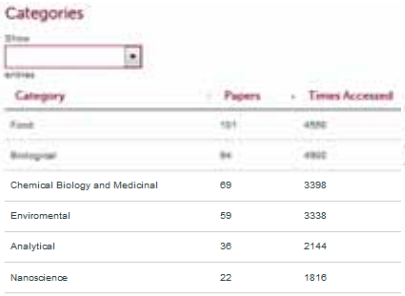
Molecular Systems Design & Engineering

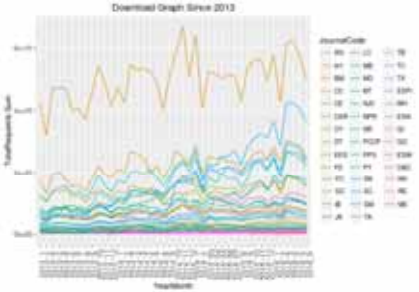
Building and designing systems from the molecular level

Now indexed in Scopus & the Web of Science Emerging Sources Citation Index (ESCI)

Any questions?
www.rsc.org/data-science







References

1. Y. Wang and L. Chen, *Nanomed.: Nanotechn*
2. A. P. Alivisatos, *Science*, 1996, **271**, 933–93
3. A. Priyam, D. E. Bluanling and K. L. Knappe
4. F. Guo, Y. Zhu, X. Yang and C. Li, *Mater. Ch*
5. P. Wang, Y. Zhu, X. Yang, C. Li and H. L. Du
6. Y. Li, Y. Zhu, X. Yang and C. Li, *Cryst. Grow*