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)
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 $[x_0 \ x_1 \ \ldots \ x_{3499}]$
Number of groups → [ 3, 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
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)

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution</td>
<td>0.7391</td>
<td>0.7628</td>
<td>0.7508</td>
</tr>
<tr>
<td>Dense</td>
<td>0.7402</td>
<td>0.7340</td>
<td>0.7370</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.6973</td>
<td>0.8049</td>
<td>0.7472</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.7592</td>
<td>0.6984</td>
<td>0.7275</td>
</tr>
<tr>
<td>Baseline: random forest</td>
<td>0.7233</td>
<td>0.5633</td>
<td>0.6334</td>
</tr>
<tr>
<td>Model</td>
<td>Me</td>
<td>OMe</td>
<td>CH₂</td>
</tr>
<tr>
<td>---------------</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dense</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convolution</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F₁ = Best result

0.4 0.6 0.8
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
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

... the quisqualic acid-induced increase in the intracellular calcium ion concentration ...
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

<table>
<thead>
<tr>
<th>“Traditional”</th>
<th>“Minimalist”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token level – modified Oscar tokeniser</td>
<td>Character level – no tokeniser</td>
</tr>
<tr>
<td>Rich feature set + token embeddings</td>
<td>Character embeddings only</td>
</tr>
<tr>
<td>Uses external resources (e.g ChEBI)</td>
<td>No external resources</td>
</tr>
<tr>
<td>1 recurrent layer</td>
<td>3 recurrent layers</td>
</tr>
<tr>
<td>Trains in hours</td>
<td>Trains in days</td>
</tr>
<tr>
<td>Large model file</td>
<td>Small model file</td>
</tr>
</tbody>
</table>

Features taken from:

Cascaded classifier for confidence-based chemical named entity recognition.
## Feature examples

<table>
<thead>
<tr>
<th>Feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regex</td>
<td><code>/^[a-z][a-z].*$/</code></td>
</tr>
<tr>
<td></td>
<td>`/^.<em>([A-Za-z].</em>[0-9]</td>
</tr>
<tr>
<td>In a list?</td>
<td>ChEBI, /usr/share/dict/words</td>
</tr>
<tr>
<td>character n-gram</td>
<td>thyl</td>
</tr>
<tr>
<td>Prefix or suffix?</td>
<td><code>^tri</code></td>
</tr>
<tr>
<td>Length</td>
<td>one$</td>
</tr>
</tbody>
</table>
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”
import

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

<table>
<thead>
<tr>
<th>System</th>
<th>Official F</th>
<th>Official P</th>
<th>Official R</th>
<th>Internal F</th>
<th>Internal P</th>
<th>Internal R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>0.8919</td>
<td>0.8867</td>
<td>0.8971</td>
<td>0.8703</td>
<td>0.8648</td>
<td>0.8758</td>
</tr>
<tr>
<td>Minimalist</td>
<td>0.8901</td>
<td>0.8865</td>
<td>0.8936</td>
<td>0.8664</td>
<td>0.8479</td>
<td>0.8858</td>
</tr>
<tr>
<td>Ensemble</td>
<td><strong>0.9032</strong></td>
<td><strong>0.9002</strong></td>
<td><strong>0.9062</strong></td>
<td><strong>0.8807</strong></td>
<td><strong>0.8646</strong></td>
<td><strong>0.8976</strong></td>
</tr>
</tbody>
</table>

## BioCreative V.5 top three

<table>
<thead>
<tr>
<th>Team</th>
<th>Method</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buzhou Tang group (Shenzhen)</td>
<td>Bidirectional LSTM (word embeddings and character embeddings)</td>
<td>0.9031</td>
</tr>
<tr>
<td>Zhihao Yang group (Dalian)</td>
<td>Bidirectional LSTM plus CRF (word embeddings and character embeddings)</td>
<td>0.9042</td>
</tr>
<tr>
<td>RSC</td>
<td>Bidirectional LSTM (ensemble of word embeddings and character embeddings)</td>
<td>0.9032</td>
</tr>
</tbody>
</table>
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!
<table>
<thead>
<tr>
<th>CPR</th>
<th>Y</th>
<th>CHEMICAL</th>
<th>GENE-N</th>
<th>Arg1</th>
<th>Arg2</th>
</tr>
</thead>
<tbody>
<tr>
<td>15110853</td>
<td>T11</td>
<td>methazolamide</td>
<td>599</td>
<td>772</td>
<td></td>
</tr>
<tr>
<td>15110853</td>
<td>T14</td>
<td>human CA</td>
<td>607</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1020</td>
</tr>
<tr>
<td>Development</td>
<td>612</td>
</tr>
<tr>
<td>Test</td>
<td>3399</td>
</tr>
</tbody>
</table>

Your model here

- No relation
- Up-regulator/activator
- Down-regulator/inhibitor
- Agonist
- Antagonist
- Substrate
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.

<table>
<thead>
<tr>
<th>INPUT</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 1 1 1 ...</td>
<td>...</td>
</tr>
<tr>
<td>the against used of magnitude ...</td>
<td>...</td>
</tr>
<tr>
<td>of the same order of magnitude</td>
<td>...</td>
</tr>
<tr>
<td>... of the against used of</td>
<td>...</td>
</tr>
<tr>
<td>... 1 1 0 0 1</td>
<td></td>
</tr>
</tbody>
</table>
Pretraining network
Recognition network
import

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Precision</th>
<th>Recall</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development</td>
<td>0.5652</td>
<td>0.7042</td>
<td>0.6271</td>
</tr>
<tr>
<td>Competition entry</td>
<td>0.5610</td>
<td>0.6784</td>
<td>0.6141</td>
</tr>
</tbody>
</table>
## Ablation studies

<table>
<thead>
<tr>
<th>Run</th>
<th>Precision</th>
<th>Recall</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random embeddings</td>
<td>0.4505</td>
<td>0.5066</td>
<td>0.4770</td>
</tr>
<tr>
<td>Public GloVe embeddings</td>
<td>0.6169</td>
<td>0.5696</td>
<td>0.5923</td>
</tr>
<tr>
<td>Chemically-trained GloVe embeddings</td>
<td>0.6297</td>
<td>0.5725</td>
<td>0.5997</td>
</tr>
<tr>
<td>Chemically-trained GloVe embeddings + transfer learning</td>
<td>0.5652</td>
<td>0.7042</td>
<td>0.6271</td>
</tr>
<tr>
<td>Team</td>
<td># runs</td>
<td>Best precision</td>
<td>Best recall</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Peng (NCBI, NLM, NIH)</td>
<td>5</td>
<td>0.7437</td>
<td>0.5735</td>
</tr>
<tr>
<td>Corbett (RSC)</td>
<td>1</td>
<td>0.5610</td>
<td>0.6784</td>
</tr>
<tr>
<td>Mehryary (Turku)</td>
<td>3</td>
<td>0.6608</td>
<td>0.6006</td>
</tr>
<tr>
<td>Lim (Korea University)</td>
<td>2</td>
<td>0.6760</td>
<td>0.5194</td>
</tr>
<tr>
<td>Lung (Florida State)</td>
<td>2</td>
<td>0.6352</td>
<td>0.5121</td>
</tr>
</tbody>
</table>

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.
Chemlistem - chemical named entity recognition using recurrent neural networks.
Full writeup: *J. Cheminf.*, submitted.

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

Editorial Board

Juan de Pablo (Chair)
Institute for Molecular Engineering, University of Chicago, USA

Claire Adjiman
Imperial College London, UK

David Awschalom
Institute for Molecular Engineering, University of Chicago, USA

Samson Jenekhe
University of Washington, USA

Kristi Kiick
University of Delaware, USA

Yongye Liang
Southern University of Science and Technology, China

Andrew de Mello
ETH Zürich, Switzerland

Marcus Müller
Institute for Theoretical Physics, University of Göttingen, Germany

Niren Murthy
University of California, Berkeley, USA

Paul Raithby
University of Bath, UK

Yuriy Román
MIT, Cambridge, USA

Now indexed in Scopus & the Web of Science Emerging Sources Citation Index (ESCI)
Any questions?
www.rsc.org/data-science