
# CheMU: ChemIE on Patents using Deep Learning

**Saber Akhondi** **Camilo Thorne** **Zenan Zhai** **Dat Quoc Nguyen**
**Christian Druckenbrodt** **Karin Verspoor** **Trevor Cohn**
Elsevier Life Sciences The University of Melbourne

## Overview

- Research project run by Elsevier (50%) and Prof. K. Verpoor's bioNLP group at the University of Melbourne (50%)
- 3-year project (Feb 2019 - Jan 2021) with yearly reviews
- Team: 1 PostDoc (100%), 2PhD students (100%) & 2 professors (Melbourne); 2 data scientists and 1-2 chemists (Elsevier)

## Objectives

Study a number of open NLP research problems in the domain of chemistry patents:

**Q1** Can we leverage deep learning-based NER models to build models that detect chemical compounds in patents?

**Q2** Can we build reaction extraction models that detect and resolve reactions and their reactants in patents?

**Q3** Can we build models that detect compounds and extract relations from tables?

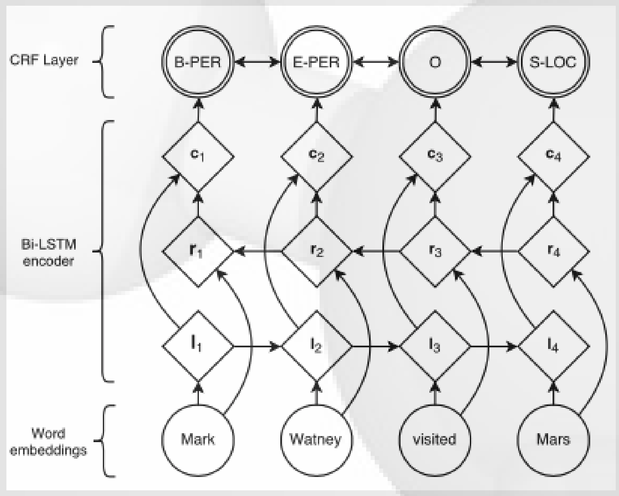
**Q4** Can we leverage high quality Reaxys patent data to achieve the state-of-the-art in these three tasks?

**Q5** Relesase datasets and organize shared ChemIE tasks to engage the academic and R&D community

## Publications

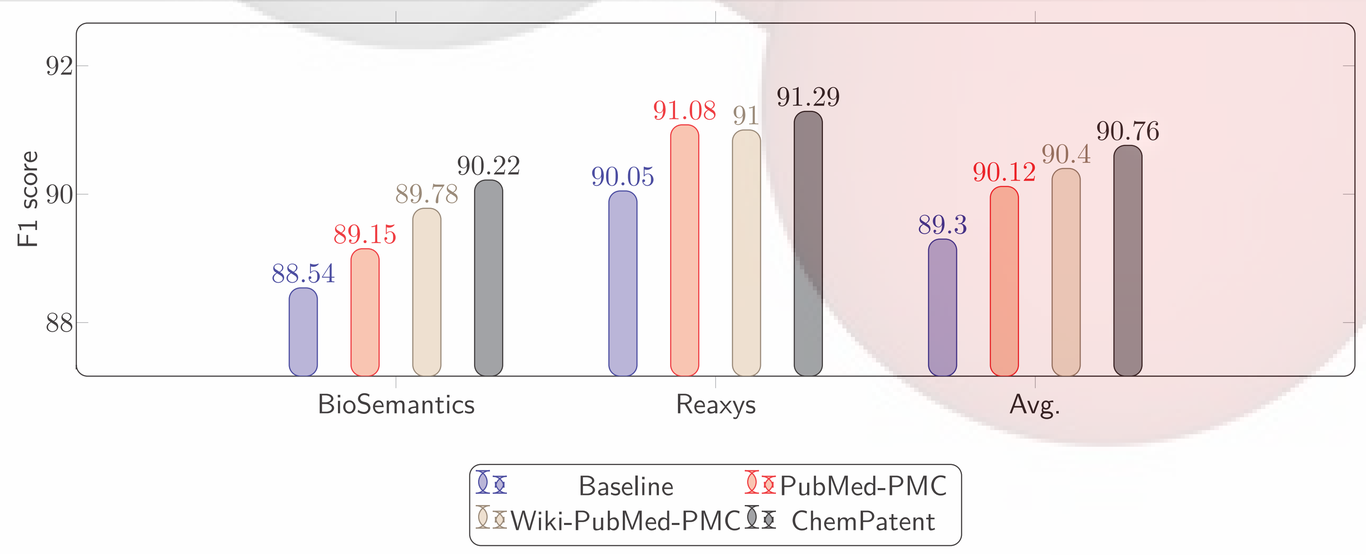
[ZZ19] Saber Akhondi Camilo Thorne Christian Druckenbrodt Trevor Cohn Michelle Gregory Karin Verspoor Zenan Zhai, Dat Quoc Nguyen. **Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings**. In *Proceedings BioNLP@ACL2019*, 2019.

## Chemical Compound Detection (Done)




- **Task:** find compounds in WO, US, IN, GB, AU, EP, CA patents
- **Dataset:** Reaxys NER GS + patents
- **Method:** ELMo 1B word embedding + biLSTM-CRF NER model
- **Results:** best model known (2019), close to human performance

| Entity label | Distr. | BiLSTM-CNN-CRF | | | +ELMo | | |
|---|---|---|---|---|---|---|---|
| | % | P | R | $F_1$ | P | R | $F_1$ |
| Class | 12.36 | 78.35 | 66.46 | 71.92 | 81.96 | 75.75 | 78.73 |
| $\text{Class}_{\text{biomol}}$ | 7.96 | 71.86 | 70.50 | 71.17 | 76.27 | 78.76 | 77.50 |
| $\text{Class}_{\text{markush}}$ | 0.32 | 42.86 | 47.37 | 45.00 | 42.86 | 47.37 | 45.00 |
| $\text{Class}_{\text{mix}}$ | 3.24 | 76.49 | 59.69 | 67.05 | 74.18 | 64.60 | 69.06 |
| $\text{Class}_{\text{mix-part}}$ | 1.35 | 71.00 | 44.10 | 54.41 | 78.10 | 50.93 | 61.65 |
| $\text{Class}_{\text{poly}}$ | 5.10 | 81.40 | 72.82 | 76.87 | 89.20 | 84.07 | 86.56 |
| Comp | 58.53 | 89.02 | 92.01 | 90.49 | 91.01 | 94.58 | 92.76 |
| $\text{Comp}_{\text{mix-part}}$ | 7.57 | 90.02 | 81.86 | 85.75 | 90.63 | 85.62 | 88.05 |
| $\text{Comp}_{\text{proph}}$ | 3.57 | 18.52 | 2.35 | 4.17 | 77.75 | 79.58 | 78.65 |
| **Micro Avg.** | 100.0 | 85.12 | 80.36 | 82.67 | 87.41 | 87.53 | 87.47 |




## Table Extraction (Ongoing)

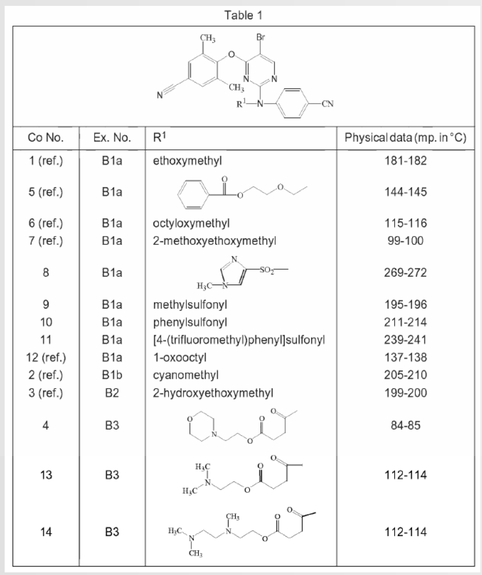


Table 1

| Co No. | Ex. No. | $R^1$ | Physical data (mp. in °C) |
|---|---|---|---|
| 1 (ref.) | B1a | ethoxymethyl | 181-182 |
| 5 (ref.) | B1a |  | 144-145 |
| 6 (ref.) | B1a | octyloxymethyl | 115-116 |
| 7 (ref.) | B1a | 2-methoxyethoxymethyl | 99-100 |
| 8 | B1a |  | 269-272 |
| 9 | B1a | methylsulfonyl | 195-196 |
| 10 | B1a | phenylsulfonyl | 211-214 |
| 11 | B1a | [4-(trifluoromethyl)phenyl]sulfonyl | 239-241 |
| 12 (ref.) | B1a | 1-oxooctyl | 137-138 |
| 2 (ref.) | B1b | cyanomethyl | 205-210 |
| 3 (ref.) | B2 | 2-hydroxyethoxymethyl | 199-200 |
| 4 | B3 |  | 84-85 |
| 13 | B3 |  | 112-114 |
| 14 | B3 |  | 112-114 |

- **Task:** classify table into Reaxys excerption guidelines categories
- **Dataset:** 7,886 tables annotated with Reaxys content types (compounds, reactions, spectroscopic data, etc.)
- **Challenge:** meaning = text + table structure
- **Proposal:** learn table embedding and apply image classification (e.g. PixelRNN, ResNet)