CheMU: bioChemical text Mining for advancing chemical and pharmaceUtical knowledge

Camilo Thorne

Elsevier Content Transformation, Life Sciences

Amsterdam, September 2, 2019

Overview

- Research project funded by Elsevier (50%) and the bioNLP group at the University of Melbourne (50%)
- **Budget**: approx. 1.8M US$
- **Duration**: 3-years (2018 – 2021), with yearly reviews
- **Core team**:
  - ▷ 1 PostDoc (100%), 2 PhDs (100%), 2 Profs (Uni Melbourne)
  - ▷ 2 data scientists, 2 chemists (Elsevier)

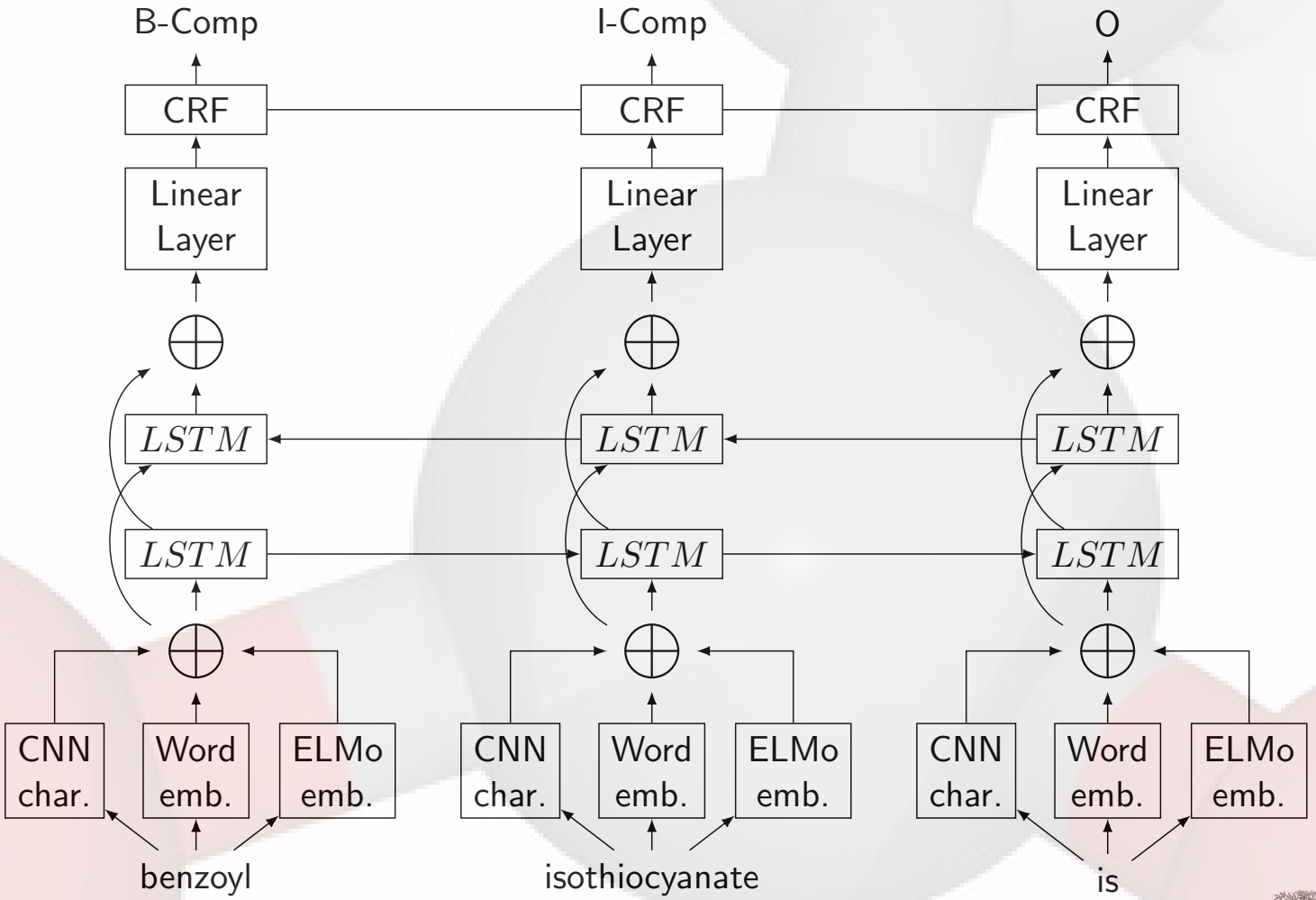
Objectives

1. Can we leverage DL NER models to **detect chemical compounds** in patents?
2. Can we build **reaction extraction** models for patents?
3. Can we run such models over patent **tables**?
4. Can we leverage **Reaxys®data** to achieve the state-of-the-art?
5. Release datasets and organize **shared tasks** to engage the academic and R&D community

Chemical NER

- **Task**: detect chemical compounds in patents
- **Datasets**:
  - ▷ Reaxys®chemical NER gold standard
  - ▷ Biosemantics chemical NER gold standard
  - ▷ 1B corpus of US, EP, WO, IN, GB, AU, CA patents
- **Methods**:
  - ▷ ELMo word embedding
  - ▷ biLSTM-CRF NER model
- **Results**: best model known (2019) for this task

Chemical NER - biLSTM-CRF Model



Chemical NER - Results

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

CLEF Shared Task

- **Task:** organize shared IE tasks over patents, viz.,
  - ▷ NER for chemical compounds
  - ▷ reaction extraction
- **Gold standard creation:**
  - ▷ 1,5K reaction snippets from patents (∼7K sentences)
  - ▷ fine-grained annotation of compounds and reactions
- **Status:**
  accepted at CLEF 2020 in Thessaloniki

ELSEVIER

THE UNIVERSITY OF MELBOURNE

CLEF Shared Task

10.0 g (35.0 mmol) of **2-tert-butyl 4-ethyl 5-amino-3-methylthiophene-2,4-dicarboxylate** (Example 1A) were dissolved in 500 ml of **dichloromethane** and 11.4 g (70.1 mmol) of **N,N'-carbonyldiimidazole** (CDI) and 19.6 ml (140 mmol) of **triethylamine** were added.

ID	Entity type	Text span
T1	Starting_material	2-tert-butyl 4-ethyl 5-amino-3-methylthiophene-2,4-dicarboxylate
T2	Solvent	dichloromethane
T3	Starting_material	N,N'-carbonyldiimidazole
T4	Reagent	triethylamine
T5	Reaction step trigger	dissolved
T6	Reaction step trigger	added

ID	Event type	Event trigger	Argument_1	Argument_2	Argument_3
E1	Reaction step	T5	Theme:T1	Theme:T2	
E2	Reaction step	T6	Theme:E1	Theme:T3	Theme:T4

Expanding scope from NER to reactions and more!

Current Project Outcomes

1. **NER**: Zenan Zhai, Dat Quoc Nguyen, Saber A. Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory and Karin Verspoor. "Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings". In: *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019*. https://www.aclweb.org/anthology/W19-5035/
2. **Shared task**: Shared task/Lab proposal accepted at the 11th CLEF (Cross Language Evaluation Forum) Conference (Thessaloniki, Sep 22-25 2020). https://clef2020.clef-initiative.eu/

All models, preprocessing scripts and datasets archived in Elsevier!

Thanks!

(Elsevier)		(Uni Melbourne)	
Saber Akhondi - PI	(Data Scientist)	**Karin Verspoor - PI**	(Prof.)
Camilo Thorne	(Data Scientist)	Trevor Cohn	(Prof.)
Christian Druckenbrodt	(Chemist)	Dat Quoc Nguyen	(PostDoc)
Ralph Hössel	(Chemist)	Zenan Zhai	(PhD)
		Biaoyan Fang	(PhD)