ELSEVIER

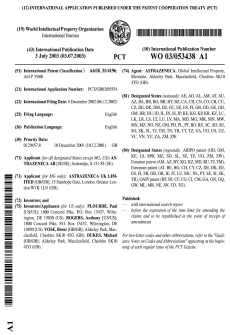
NLP4Chemistry

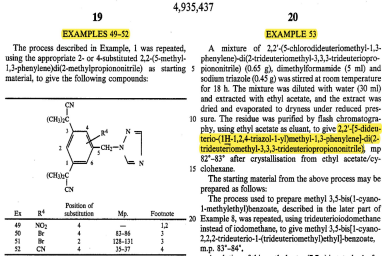
Text Mining Applications

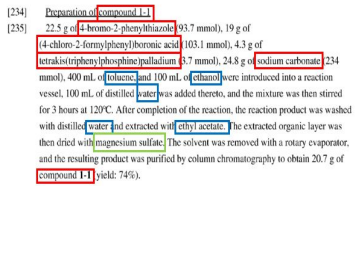
Camilo Thorne, Saber Akhondi

May 2024 - COLING/LREC '24

Chemical information extraction







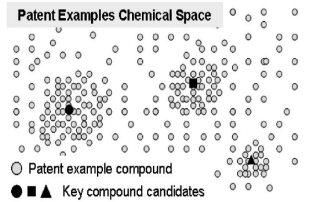


Figure 2. Graphical image of patent example compounds in chemical space. Each gray circle represents an example compound. The black circle, square, and triangle represent key compound candidates.

Theory: Chemists carry out extensive SAR around key compounds. Cluster examples and look for centres of densely populated regions

Read relevant document

Look for compounds

Focus on relevant compounds

Identify role in patent and reaction

Predict key compounds

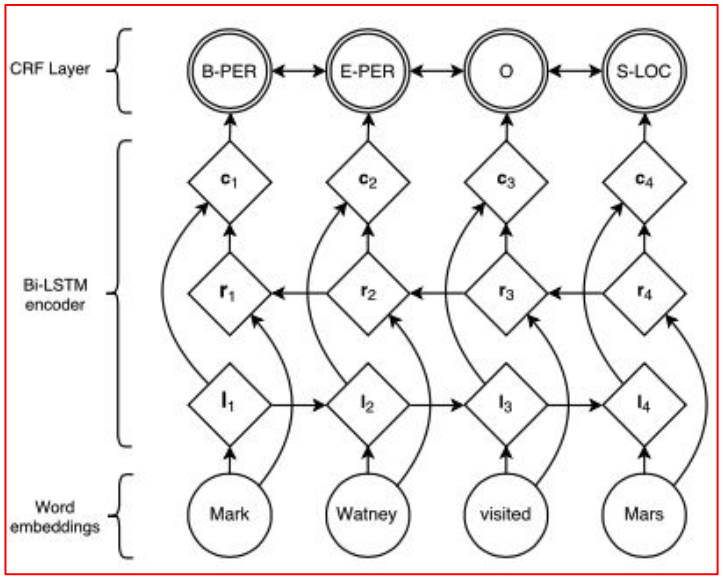
Key NLP tasks

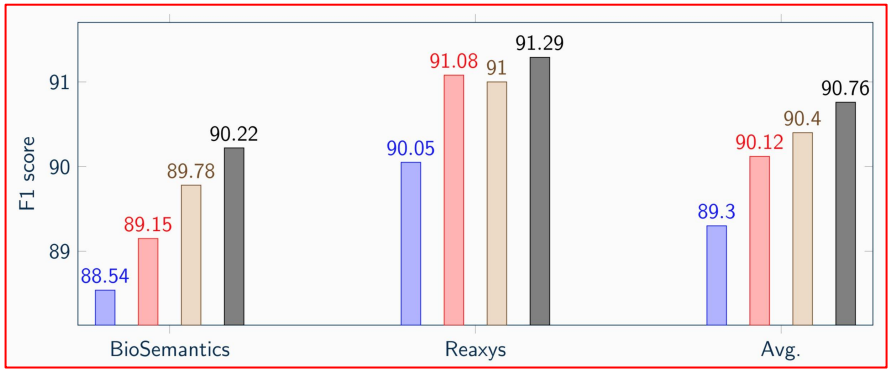
- NER: **Named entity recognition (chemical entities)**
  - Identify chemical compounds
  - Detect the role they play in chemical reactions, e.g. ***Starting_material*** or ***Solvent***
- EE/RE: **Event extraction (chemical reactions)**
  - Identify a sequence of event steps of a chemical reaction
  - Involves event trigger detection, event typing and thematic role recognition

Chemical NER

Chemical NER on patents [Zhai, 2019; He 2021]

- CRFs (2014), BiLSTM-CRFs (2019), BERT since





Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings

Authors: Zenan Zhai, Dat Quoc Nguyen, Saber A Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, Karin Verspoor

Publication date: 2019/7/5

Definitions of entities [He, 2021]

Label	Definition
REACTION_PRODUCT	A substance that is formed during a chemical reaction.
STARTING_MATERIAL	A substance that is consumed in the course of a chemical reaction providing atoms to products.
REAGENT_CATALYST	A compound added to a system to cause or help with a chemical reaction. Compounds like catalysts, bases to remove protons or acids to add protons must be also annotated with this tag.
SOLVENT	A chemical entity that dissolves a solute resulting in a solution.
OTHER_COMPOUND	Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents.
EXAMPLE_LABEL	A label associated with a reaction specification.
TEMPERATURE	The temperature at which the reaction was carried out.
TIME	The reaction time of the reaction.
YIELD_PERCENT	Yield given in percent values.
YIELD_OTHER	Yields provided in other units than %.

Reaction Extraction

Reaction Extraction From Patents [He, 2021]

- The chemical and pharmaceutical industries depend on the discovery of new chemical compounds
- Most new compounds and their synthesis are described only in **patent documents**
- Patents too abundant for manual processing
- NLP approaches can enable **automatic reaction extraction** from chemical patents and support compound discovery and synthesis

❑ **Problems:**

1. no gold standards
2. *patents are written very differently as compared to scientific literature*
3. task(s) ill-understood

Chemical Reaction Example [He, 2021]

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	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	Trigger	dissolved
T6	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

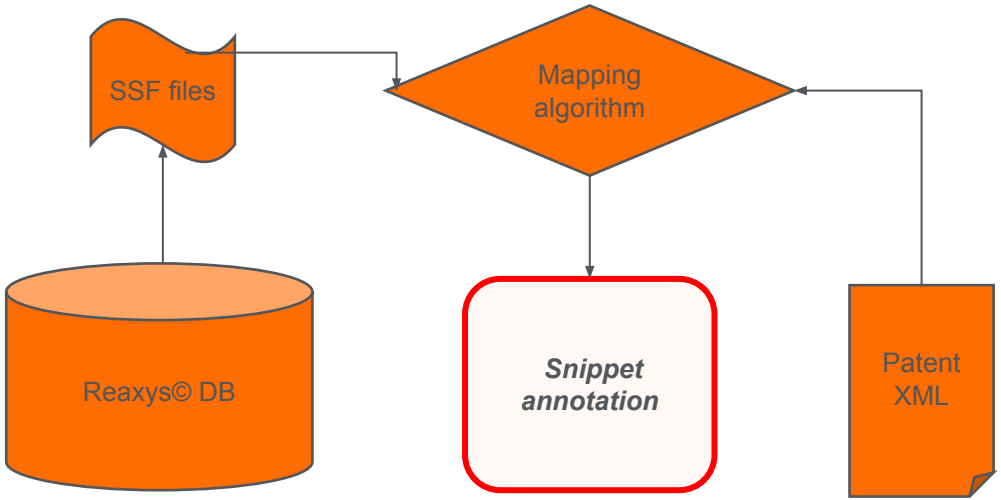
Task 1 – NER – in Red
Task 2 – Event extraction – in Purple

Definitions of Events [He, 2021]

Label	Definition
WORKUP	Within a WORKUP event, the chemical product is only isolated, i.e. this event type refers to the series of manipulations required to isolate and purify the product(s) of a chemical reaction.
REACTION_STEP	Within a REACTION_STEP event, the starting materials are converted into the product.
Arg1	Arg1 represents **argument roles** of being causally affected by another participant in the events. It labels the relation between an event trigger word and a **chemical compound**.
ArgM	ArgM represents adjunct roles with respect to an event. It labels the relation between a trigger words and a **temperature, time, or yield entity**.

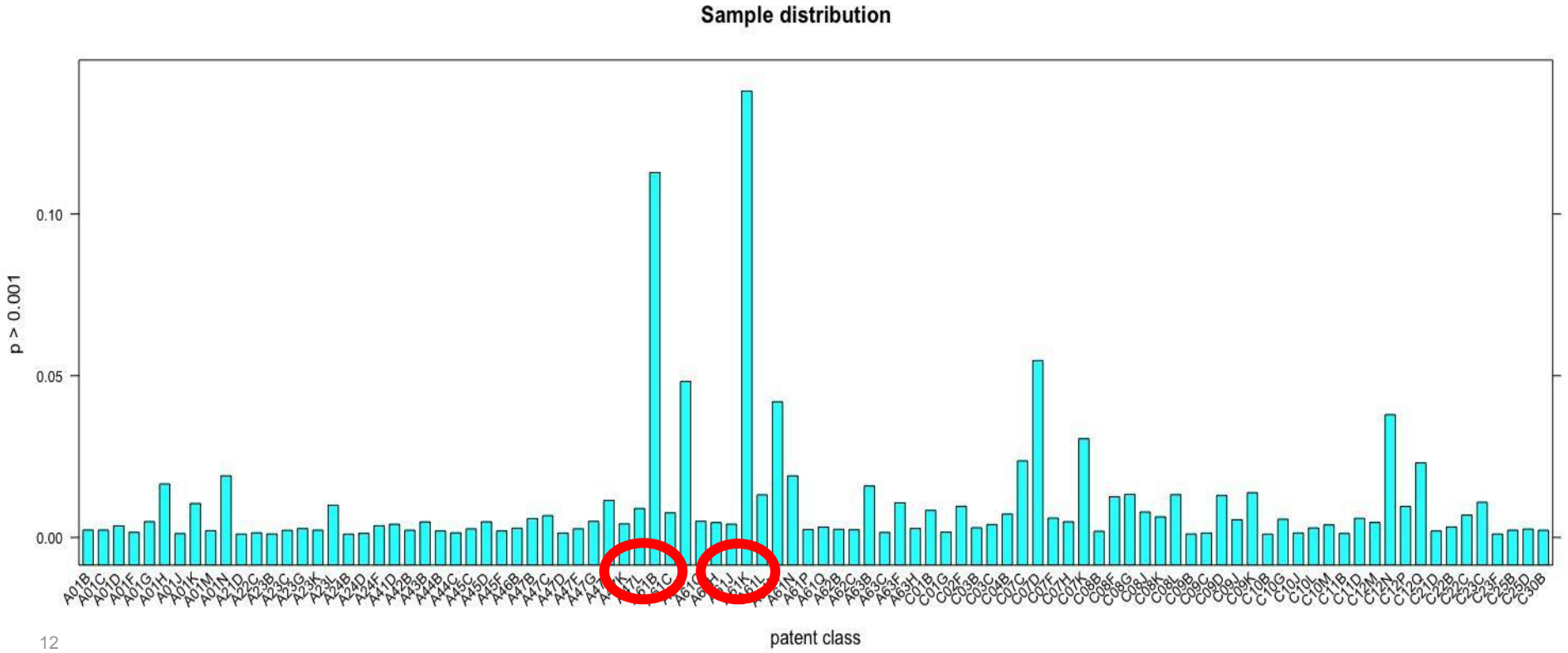
Creation of the dataset [He, 2021]

- **STEP 1:** Mapping of excerpted data from Reaxys© DB to Patent XML files (pre-annotation)
- **STEP 2**: 3,000 pre-annotated snippets selected for gold annotation (7,000 sentences)
- **STEP 3:** Each snippet annotated (entities + events) by: 2 independent annotators and 1 harmonizer

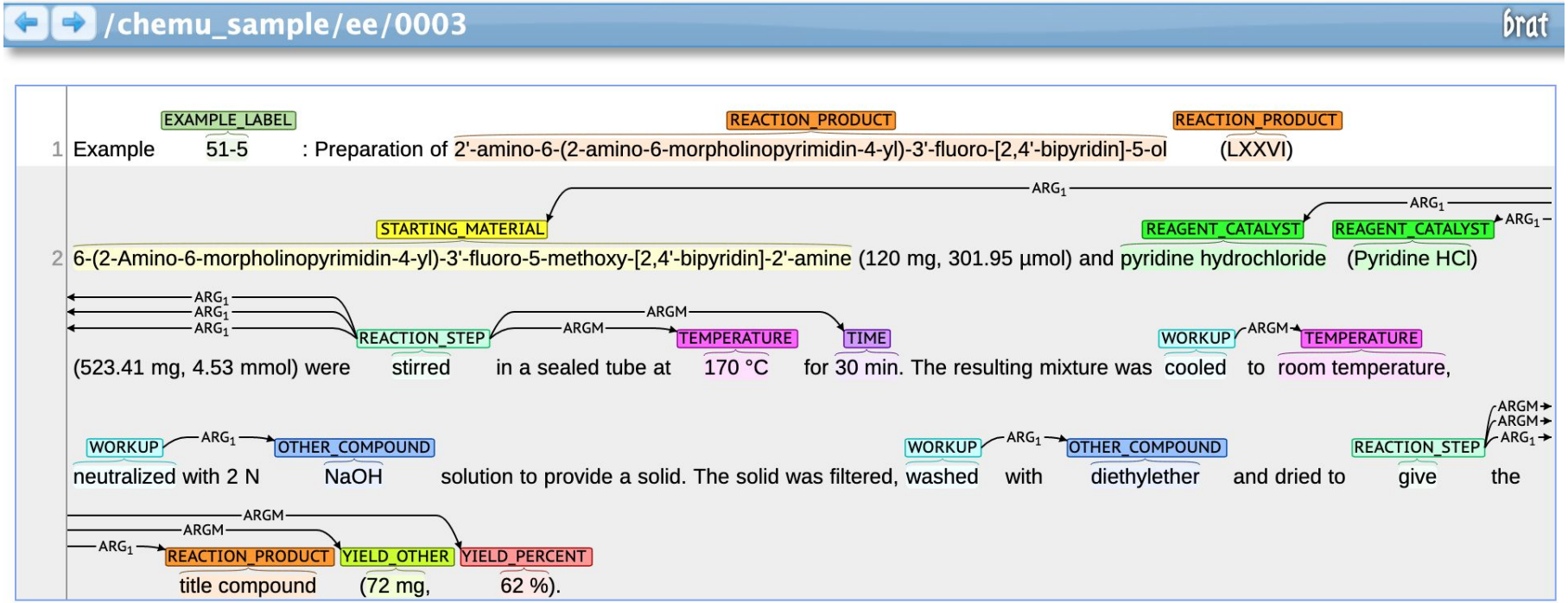


Creation of the dataset [He, 2021]

- **Distribution**: 6 patent offices (AU, US, EP, IN, AU, WO), with focus on **organic chemistry** (> 15%)



BRAT Format [He, 2021]



Evaluation

- Standard metrics of Precision / Recall / F-score
  1) Exact matching
  2) Approximate span matching

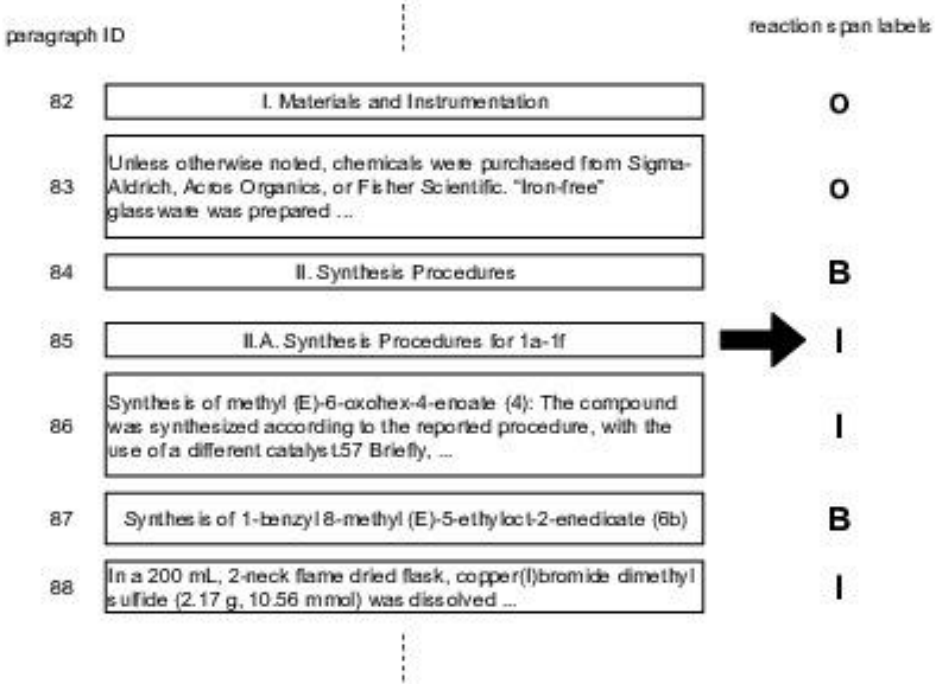
- Three settings:
  1) NER: detect and classify reactants
  2) Event extraction (EE): event extraction only, given entities
  3) End-to-end: NER + EE in one go

Results [He 2021, F1-score]

Team	NER	NER (a)	EE	EE(a)	E2E	E2E (a)
Melax	**0.98**	**0.99**	**0.95**	**0.95**	**0.92**	**0.93**
NextMove	0.93	0.97	0.90	0.90	0.80	0.82
OntoChem	0.78	0.82	--	--	0.51	0.54
BOUN_REX	--	--	0.72	0.72	--	--
NLP@VCU	0.89	0.98	0.65	0.65	--	--
AUKBC	0.57	0.74	--	--	--	--
Lassige_BioTM	0.96	0.99	--	--	--	--
KFU_NLP	0.86	0.97	--	--	--	--
VinAI	0.96	0.99	--	--	--	--
SSN_NLP	0.25	0.68	--	--	--	--
BiTeM	0.94	0.98	--	--	--	--
JU_INDIA	0.15	0.42	--	--	--	--

Challenges

Reaction span detection [Yoshikawa, 2021]

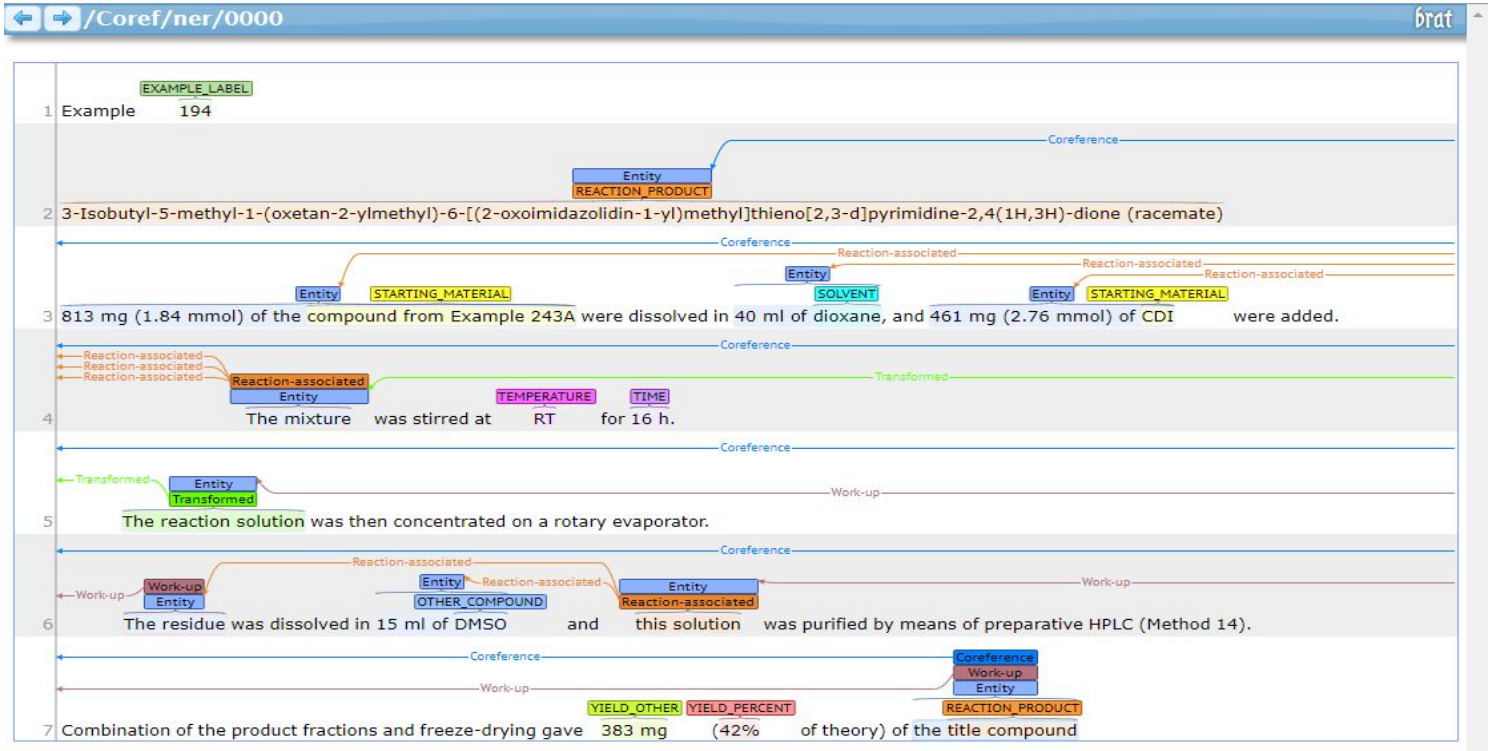


Reaction span detection [Yoshikawa, 2021]

Decoder	Input token representation	Strict match			Fuzzy match		
		$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
Rule-based		.205	.381	.241	.278	.482	.319
Logistic		.421	.380	.376	.521	.462	.461
Paragraph-level softmax	w2v +ELMo	.352	.365	.336	.475	.457	.437
	w2v +ELMo +$\text{NER}_{\text{COARSE}}$	.340	.389	.337	.446	.468	.415
	w2v +ELMo +NER_{FINE}	.345	.383	.341	.479	.485	.447
Paragraph-trigram softmax	w2v +ELMo +NER_{FINE}	.513	.488	.482	.643	.573	.574
BiLSTM-CRF	w2v +ELMo +NER_{FINE}	**.658**	**.653**	**.640**	**.718**	**.708**	**.696**

Anaphora resolution [Fang, 2022]

Resolve anaphors within and across reaction snippets to build full reaction:



Relevant compounds in journals and patents [Akhondi, 2019]

Substances from text

1. Context
2. Section – Title, Abstract, References..etc.
3. Correctness vs Relevancy

   - Reagents, Solvents, Catalyst etc.

   ……

Challenge: Which of the substances extracted from the article are **relevant**?

1

of this iron-catalyzed hydroaminocarbonylation. As shown in Scheme 3, the reactions of aromatic internal alkynes (**3a-3e**) afforded the corresponding products in high yields (84–89%). For aromatic internal alkynes (**3f**), the yield decreased, affording the corresponding succinimide in moderate yield (41%), steric reasons are likely to be responsible for the moderate yield. Moreover, it was shown that our attempt failed when phenylacetylene **3g** was employed as substrate.

2

[25] P. Riviere, K. Koga, An approach to catalytic enantioselective protonation of prochiral lithium enolates, Tetrahedron Lett. 38 (1997) 7589–7592, https://doi.org/10.1016/S0040-4039(97)01790-5.

[26] J. Eames, N. Weerasooriya, Investigations into the enantioselective protonation of enolates derived from 2-methyl-1-tetralone using a chiral diamine ligand, Tetrahedron Lett. 41 (2000) 521–523, https://doi.org/10.1016/S0040-4039(99)02109-7.

[27] T. Poisson, V. Dalla, F. Marsais, G. Dupas, S. Oudeyer, V. Levacher, Organocatalytic enantioselective protonation of silyl enolates mediated by cinchona alkaloids and a latent source of HF, Angew. Chem. Int. Ed. 46 (2007) 7090–7093, https://doi.org/10.1002/anie.200701683.

3

2. Result and discussion

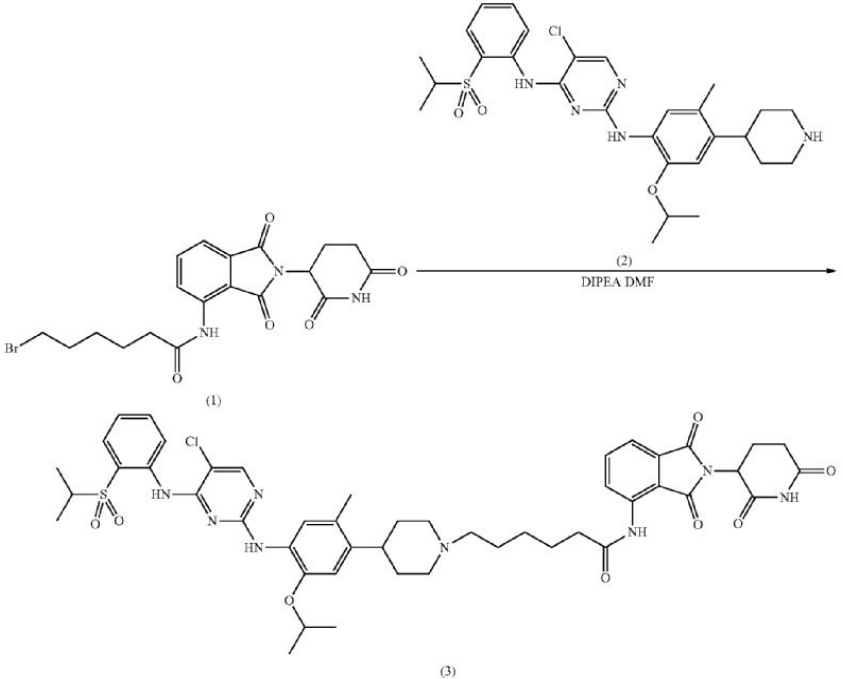
The optimization of this reaction was carried out by employing the 2-formylphenyl trifluoromethanesulfonate **1a** and benzaldehyde **2a** as substrates under the nickel catalysis (Table 1). To our delight, it exhibited that $Ni(cod)_2$ could facilitate this transformation with 74% yield by the use of triphos as ligand and 1-methylpiperidine as base in toluene (entry 1, Table 1). Encour-

Multimodality [ARC Proposal, 2024]

- Multimodality
- Resolution across text & images

[0166] Step 2:

[0167] 250 mg of compound (1), 365 mg of compound (2) and 650 mg of diisopropylethylamine were dissolved in 5 ml of N, N-dimethylformamide. The mixture was stirred at 80° C. for 6 h, and then cooled to room temperature. After concentration, the residue was purified by column chromatography to obtain 350 mg of compound (3) with a yield of 68.0%.MS (ESI): 927 $[M+H]^+$. ^1H NMR (400 MHz, $CDCl_3$) δ 9.49(s, 1H), 8.28(d, J=8.4 Hz, 1H), 8.15(s, 1H), 7.98(s, 1H), 7.95(dd, J=8.0, 1.6 Hz, 1H), 7.87(d, J=7.2 Hz, 1H),7.



Representation Learning

Word embeddings for chemistry [Thorne, 2020]

- Word embeddings trained on chemical texts can acquire substantial amounts of domain knowledge
- Validated extrinsecally via NER
- The larger and more contextual the embedding, the better the F1-score

Table 1: Our training and test sets come from the SCAI corpus; the validation set from the Biosemantics corpus.

Split	Entities	Tokens
Train	731 IUPAC, 212 Modifier, 73 Partiupac	33,457
Validation	240 IUPAC	4,654
Test	48 IUPAC, 2 Modifier	28,240

Table 2: Overview of the embeddings studied in this paper.

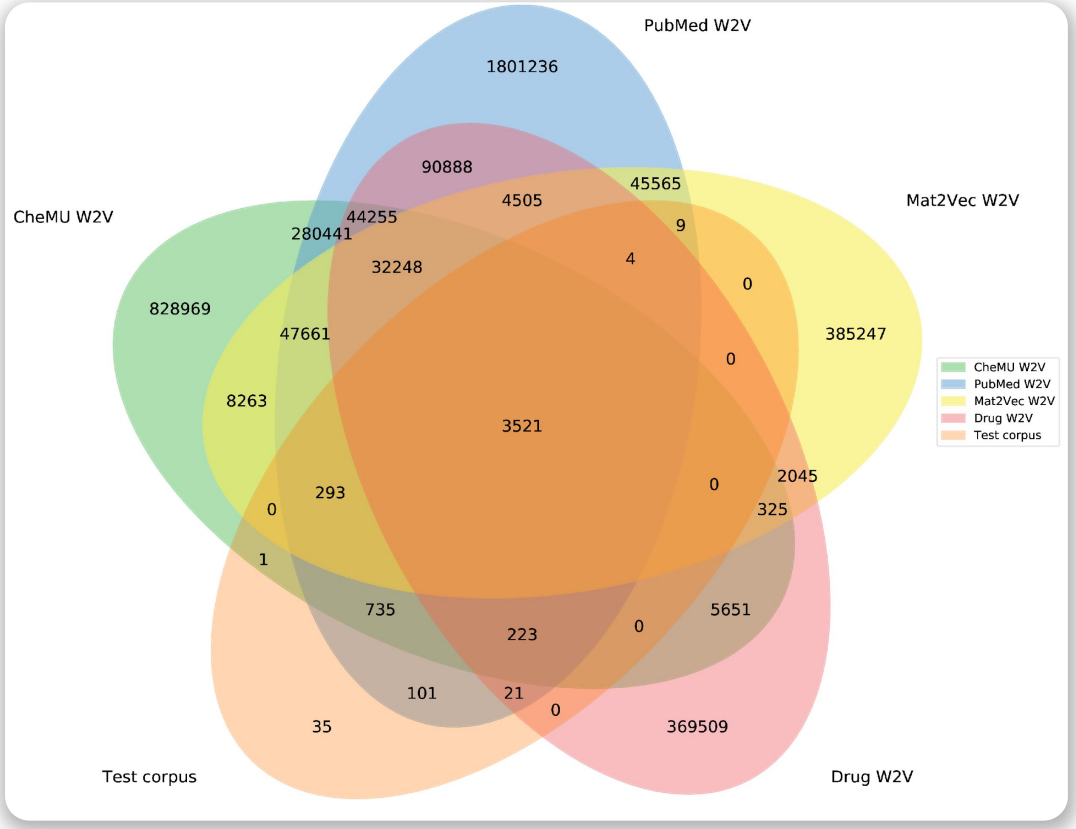
Embedding	Words	Dimensions
Mat2Vec W2V	529,686	200
PubMed W2V	2,351,706	200
Drug W2V	553,195	420
CheMU W2V	1,252,586	200
PubMed ELMo	—	1,204
CheMU ELMo	—	1,204

Table 3: Impact of the different chemical embeddings on chemical NER (sorted by F1 score).

Word Embedding	F1	Δ (F1)
Mat2Vec W2V	26.89%	—
PubMed W2V	27.23%	+ 0.3%
Drug W2V	48.48%	+21.3%
CheMU W2V	53.24%	+ 4.8%
PubMed ELMo	70.15%	+16.9%
CheMU ELMo	72.41%	+ 2.3%

Vocabulary overlap [Thorne, 2020]

- The embeddings were trained on different corpora
- Vocabularies distinct
- They overlap on 3521 content words included in the IUPAC test corpus
- We retricted all embeddings to this vocabulary and compared ensuing vectors

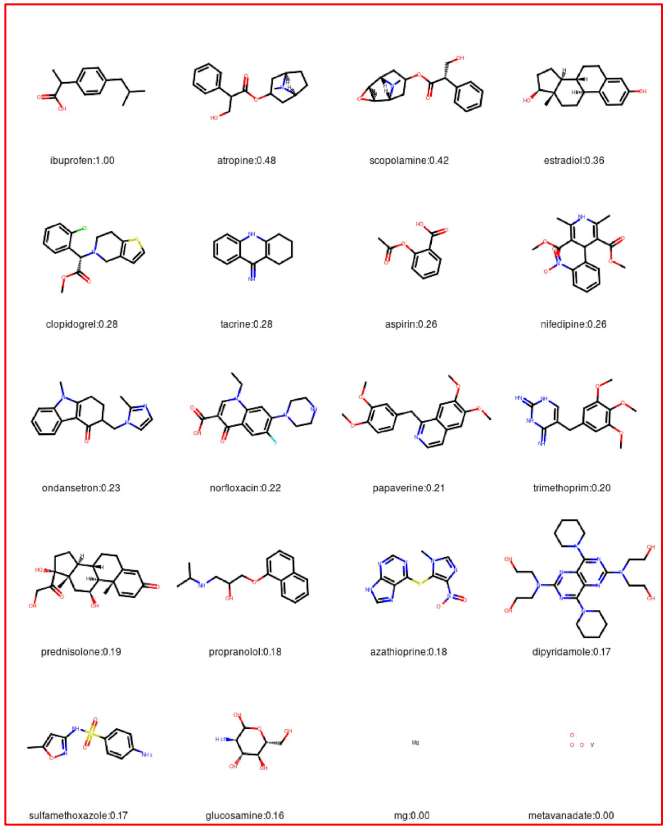


Top 20 most similar drugs to **ibuprofen** [Thorne, 2020]

- Build embeddings for "ibuprofen"
- Check for synonyms on IUPAC corpus
- Generate SMILES and check for chemical similarity (fingerprint)
- Semantic similarity aligns with structural similarity

Table 4: Top 10 similarity lists ("ibuprofen" query).

CheMU ELMo	PubMed ELMo	CheMU W2V	PubMed W2V	Drug W2V	Mat2Vec W2V
tacrine	atropine	aspirin	aspirin	pronounced	drug
ondansetron	ondansetron	clopidogrel	ondansetron	ultrastructure	drugs
aspirin	sulfamethoxazole	prednisolone	clopidogrel	mimics	aspirin
clopidogrel	aspirin	azathioprine	propranolol	surgical	sulfamethoxazole
dipyridamole	tacrine	atropine	placebo	favorable	propranolol
atropine	trimethoprim	nifedipine	tacrine	intestine	trimethoprim
prednisolone	propranolol	sulfamethoxazole	nifedipine	trained	norfloxacin
propranolol	prednisolone	dipyridamole	prednisolone	extinct	estradiol
trimethoprim	clopidogrel	propranolol	mg	slightly	antibiotics
nifedipine	papaverine	papaverine	topical	combination	nifedipine

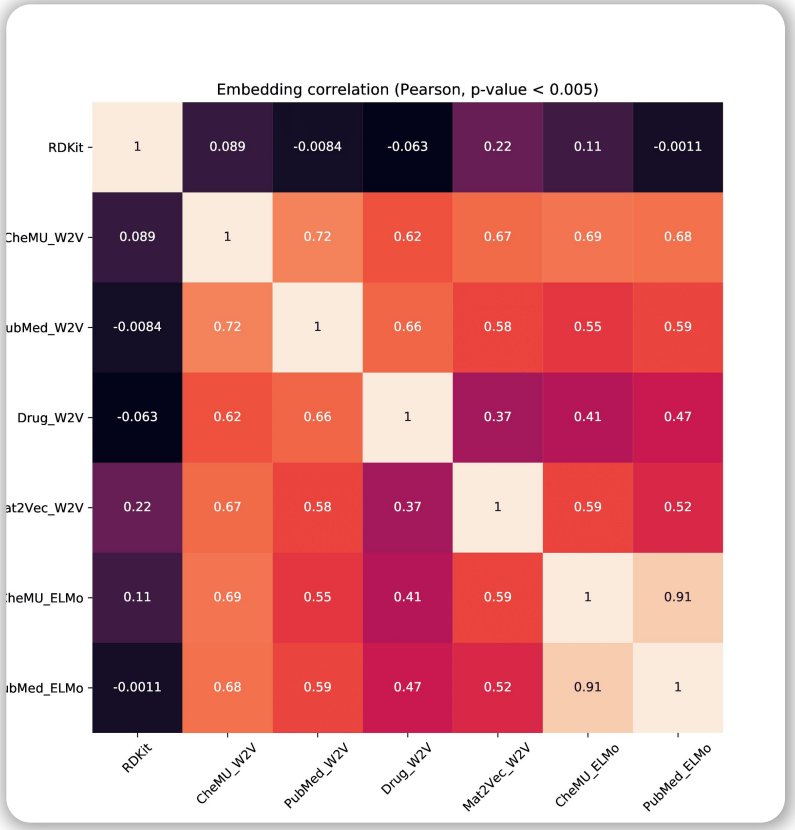


Embedding correlation analysis [Thorne, 2020]

RDKit was used to measure fingerprint similarity via SMILES

Observations

- ELMo embeddings correlate highly (0.91)
- Large W2V embeddings correlate highly (0.72)
- Material science W2V embeddings correlate positively with fingerprints (0.21)



References

- 2015 - CHEMDNER:
  The drugs and chemical names extraction challenge - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4331685/

- 2017 - Information Retrieval and Text Mining Technologies for Chemistry
  - https://pubs.acs.org/doi/abs/10.1021/acs.chemrev.6b00851

- 2020 - Word Embeddings for Chemical Patent Natural Language Processing
  - https://arxiv.org/abs/2010.12912

- 2021 - ChEMU 2020: Natural Language Processing Methods
  Are Effective for Information Extraction From Chemical Patents
  - https://www.frontiersin.org/articles/10.3389/frma.2021.654438/full

- 2023 - ChemNLP: A Natural
  Language-Processing-Based Library for Materials Chemistry Text Data
  - https://pubs.acs.org/doi/abs/10.1021/acs.jpcc.3c03106

Thank you!