Extraction of reactions from patents using grammars

Daniel Lowe[1][0000-0002-0045-5721] and John Mayfield[2][0000-0001-7730-2646]

[1] Minesoft, Cambridge, United Kingdom
[2] NextMove Software, Cambridge, United Kingdom
john@nextmovesoftware.com

Abstract. An approach to entity recognition and event annotation in synthetic chemistry text, by recognising such text using grammars, is described. LeadMine is used to recognize chemicals and physical quantities using a grammar. These entities are used with ChemicalTagger's phrase grammar to determine the relationship between chemicals and reaction properties. Finally chemical structure information is used to assign chemical role information, by inspection of the individual compounds and through whole reaction analysis techniques like NameRxn and atom-atom mapping. Our approach obtained an F1-score of 0.898 for both the named entity recognition and event annotation tasks.

Keywords: Reaction Extraction, ChemicalTagger, LeadMine, NameRxn, Grammars

1 Introduction

The extraction of reaction content from unstructured text plays an important part in the population of the reaction databases that expedite the work of synthetic chemists. The increasing size of both the patent and non-patent literature mean that there is a growing demand for automated solutions for extracting this data, although attempts to automate this task date back to 1980s [1, 2].

ChEMU proposed two tasks, the first being a named entity recognition (NER) task, and the second being an entity relationship task. The entities in question are chemicals, with different types being assigned to chemicals having different roles in a reaction, as well as reaction properties e.g. yield. The second task was to define the relationship between reaction actions and chemicals or reaction properties. A complete description of the task is present in the task paper [3].

Our approach builds upon an open source reaction extraction tool [4, 5] in combination with NextMove Software's entity recognition tool, LeadMine [6]. Our process has already been applied to millions of patent documents with the resulting extracted reactions made available through public [7] and commercial datasets [8]. The quantity of reactions extracted and CC-Zero licensing on the public dataset has resulted in

wide use for analysis of reaction trends [9], and as a data source for retrosynthesis [10] and reaction prediction [11].

In this work we adapt our process (Fig. 1) to the tasks proposed by ChEMU.

Chloro N-arylation (1.3.7)

Name	Role	Formula	MW	Amount	Mass	Volume	Density	Yield
1-(3-Chloro-2-pyrazinyl)-2-pyrrolidinone	Product	$C_8H_8ClN_3O$	197.622 g/mol	0.506 mmol	100 mg			32.96 %
2,3-Dichloropyrazine	Reactant	$C_4H_2Cl_2N_2$	148.978 g/mol	1.69 mmol	250 mg			
pyrrolidinone	Reactant	C_4H_7NO	85.105 g/mol	1.54 mmol	131.061 mg	120 µL	1.092 g/mL	
dimethylformamide	Solvent	C_3H_7NO	73.094 g/mol	64.574 mmol	4.72 g	5 mL	0.944 g/mL	
dimethylformamide	Solvent	C_3H_7NO	73.094 g/mol	25.83 mmol	1.888 g	2 mL	0.944 g/mL	
Sodium hydride	Agent	HNa	23.998 g/mol	1.62 mmol	67 mg			

Info	
Source	US20060040918A1 [0160]

Fig. 1. Example of output of reaction extraction process

2 Background to the approach

The reaction extraction process employs ChemicalTagger [12] to annotate chemical entities, chemical properties, trigger words indicating reaction actions and assign part of speech tags. ChemicalTagger uses the rule-based tokenization from OSCAR4 [13] to transform the input into a sequence of tokens. Taggers are then run sequentially with the result from the first tagger that matched being used. These taggers are a chemical entity tagger, regular expression-based trigger word tagger and a part of speech tagger. The tags assigned to trigger words also indicate the part of speech e.g. VB-STIR is a verb associated with the action of stirring. An Antlr [14] grammar recognizes a sequence of these tags and according to the rules matched transforms this sequence into a parse tree (Fig. 2) with nesting corresponding to: sentence, phrases and “concepts”. An example concept is a MOLECULE, which contains a chemical name but also may contain one of more QUANTITY concepts within which are the tokens for each physical quantity. In a subsequent step further nesting is added to the parse tree by grouping noun and verb phrases into “action phrases” based on the tag indicating the presence of an action term. This term is typically a verb e.g. elute/eluting, but can also be a noun e.g. elution. The contents of an action phrase describe an experimental action that occurred e.g. add, stir, dissolve, yield. Compounds in dissolve phrases are assigned as solvents.

Fig. 2. Example of ChemicalTagger parse trees with action phrases

In the Patent Reaction Extraction project [5] this parse tree, in combination with chemical structure information (from chemical name to structure), and heuristics is used to assign the role of each compound. These roles are: product, reactant, solvent and catalyst. If the parse tree indicates that a chemical is involved in a workup action the chemical is ignored. The system assigns compounds as reactants if no clear indication is given to the contrary. Anaphora resolution is also used to resolve references to chemical structures defined in preceding experimental sections. Finally atom mapping is used to sanity check the results and typically reactions for which an atom mapping cannot be established are rejected. A solvent may be reclassified as a reactant to obtain a successful atom mapping. The atom mapping is used to determine the stoichiometry of the chemicals in the reaction.

A more complete description of the outlined process is available in [4], which also covers how experimental chemistry text is distinguished from other text, which is a task that is not addressed by ChEMU, where instead all input contains a chemical reaction.

2.1 Differences from open source implementation

In this work ChemicalTagger was adapted to use LeadMine for chemical entity recognition and physical quantity recognition. ChemicalTagger's tokenization is adjusted accordingly such that all LeadMine entities were treated as single tokens e.g. "50 °C". ChemicalTagger's parse tree data structure was also enhanced with references back to the source tokens allowing character offsets of entities to be easily retrieved while navigating the parse tree.

In preference to atom-mapping NextMove Software's NameRxn [15] was used to identify the chemical reaction and assign an atom mapping. NameRxn is a pattern-based reaction classifier that has high precision but lower recall than an Maximum Common Substructure based atom-mapper. If a reaction is not recognized by

NameRxn we fall-back and use the Atom-Atom Mapper from the Indigo toolkit [16]. Normally if a reaction cannot be mapped by either NameRxn or Indigo we would not consider the extraction "complete" and exclude it from the output. As this task only concerns annotating entities and events we include reactions that we normally would consider incomplete.

3 Methodology

3.1 Named Entity Recognition

Each input experimental section is provided as one or more lines of text. These were split into headings and paragraphs, by considering lines containing fewer than 200 characters to be headings until the first paragraph of the experimental section was identified. It should be noted that in actual patent XML this sort of heuristic is typically not required as headings and paragraphs are distinguished by having different tags.

Instead of using our existing procedure for extracting reaction we instead based our submission primarily on the parse tree from ChemicalTagger, with the output of our reaction extraction process used to assist in chemical role assignment. The reasoning behind this was to overcome some of the significant mismatches between the output of the reaction extraction procedure and what was required for the ChEMU tasks:

- Exact character offsets are not recorded
- When a chemical appears alongside a synonym or way of referring it, only one instance of the chemical is recorded
- Workup compounds are intentionally ignored

From ChemicalTagger's parse tree some entities such as times, temperatures and yields were directly extractable (Table 1). The role of a chemical compound was set to match those from the extracted reactions with some significant corrections to account for the different definition of catalyst used. ChEMU allows a "catalyst" to contribute non-carbon atoms to a reaction, meaning that the catalyst may in fact be consumed by the reaction. This difference was accounted for by inspecting the reaction's computed atom mapping if available, additionally any "reactant" not contributing atoms to the product was assumed to be a catalyst. Any compound identified by ChemicalTagger, but that did not appear in an extracted reaction was assumed to be a workup compound. All chemicals matching the structure of a product were assigned the role product.

The role assignment for non-products was improved by using statistics from the training data. Chemical names were converted and indexed by structure (InChI), for each structure the entity type occurrence and incidence recorded. When annotating, these statistics were then used to improve the entity type assignment based on whether a structure is always a given type (i.e. must be catalyst) or never a given type (i.e. never a catalyst). If the second case was detected the most frequent assignment was used instead.

We cannot map all PROCEDURE tags directly to the EXAMPLE_LABEL entity types since only definitions and not references are required by the annotation guidelines. Definitions normally appear in the heading while references are found in the body of the text e.g. “prepared in the same manner as Synthesis Example 5”.

Solvent mixtures are normally handled by both the Antlr grammar and more recently a dedicated grammar and resolver. The dedicated grammar will tag and resolve the terms “aqueous sodium chloride”, “aqueous hydrochloric acid”, “aqueous sodium hydrogen carbonate” as a whole to generate either a Mixfile [17] or MOLfile/SMILES with relevant Sgroups [18]. The gold standard requires these mixtures to be recognized as two separate entities and so the grammar and resolver was disabled and the term “aqueous” was matched in isolation.

Additional logic was added to the yield detection to increase recall in two situations. Percentage yields may only be loosely associated with the entities to which they refer. We handle this case by identifying a single unknown percentage quantity in the paragraph and “promoting” it to a YIELD_PERCENT. Some yields specified as an amount or mass were missed if they were not grouped with a compound by the grammar. We supplement this by searching close to known percentage yields for a mass or amount and “promoting” these to YIELD_OTHER.

Table 1. shows the relationship between ChemicalTagger’s tags, or groups of tags and the ChEMU entity types, that was empirically determined.

ChemicalTagger tag/concept	**ChEMU entity type**	**Notes**
PROCEDURE	EXAMPLE_LABEL	Number or identifier inside a PROCEDURE
Chemical name (within MOLECULE) UNNAMEDMOLECULE REFERENCETOCOMPOUND (within MOLECULE)	REACTION_PRODUCT STARTING_MATERIAL	Role determined from parent action phrase and/or structure’s role in atom-atom mapped reaction
Chemical name (within MOLECULE)	REAGENT_CATALYST SOLVENT OTHER_COMPOUND	
NN-TEMP VB-HEAT	TEMPERATURE	VB-HEAT only for the verb reflux
NN-TIME	TIME	“overnight” ignored as this isn’t annotated in gold standard
NN-CHEMPROPERTY	YIELD_PERCENT	Yields with unit Percent. Qualifiers like “about” truncated
NN-CHEMPROPERTY	YIELD_OTHER	Yields with other units. Qualifiers like

		"about" truncated
VB-ADD NN-ADD VB-CHARGE VB-DROP VB-FILL VB-TREAT VB-DISSOLVE NN-DISSOLVE VB-HEAT VB-INCREASE VB-STIR VB-YIELD VB-IRRADIATE NN-IRRADIATE VB-SUSPEND VB-SYNTHESIZE	REACTION_STEP	The training set was used to identify additional words
VB-DEGASS NN-DEGASS VB-DRY VB-EXTRACT NN-EXTRACT VB-FILTER NN-FILTER VB-PARTITION NN-PARTITION VB-PHCHANGE VB-PRECIPITATE NN-PRECIPITATE VB-PURIFY NN-PURIFY VB-QUENCH VB-WASH NN-WASH VB-DILUTE VB-COOL	WORKUP	The training set was used to identify additional words

3.2 Event Extraction

The ChemicalTagger tag was used to initially assign a trigger word as being either REACTION_STEP or WORKUP (Table 1). However if the phrase contained more workup compounds than non-workup compounds this was changed to WORKUP. Conversely in the case where more non-workup compounds were within the phrase the role was switched to REACTION_STEP, but only for VB-COOL related trigger words.

The event annotation task required annotations of two types of event be assigned: ARG_1 and ARG_M. The former is used to indicate a relationship between a trigger word and chemical compound e.g. **washed** with **water**. The latter is used to indicate a relationship between a trigger word and a temperature, time or yield e.g. **stirred** at **room temperature**. In simple terms this means relationships with chemical entities are one type, while relationships with reaction properties are another type. Assignment of relationships were achieved by associating all entities in a ChemicalTagger action phrase with the trigger word responsible for the action phrase, with the following exceptions:

- Product chemicals were not associated with workup trigger words
- If a product is expected in the phrase the relationship could not involve a workup trigger word or workup compound (OTHER_COMPOUND)
- Yield entities were not associated with workup trigger words

Precision/recall of trigger words was enhanced by using the training set to identify all trigger words that appeared as false positives or false negatives and for each in turn determining whether always recognizing that word as REACTION_STEP, WORKUP or not a trigger word, improved performance on the training set. This yielded 46 words to be classified as REACTION_STEP, 26 to be classified as WORKUP and 23 that should not be trigger words. Many of these are likely to prove useful for improving ChemicalTagger's action phrase assignment, although the specific type of action phrase these correspond to would still need to be manually assigned as ChemicalTagger classifies reaction actions into more than 2 categories.

3.3 Patent Context

As our approach depends on atom mapping to determine the chemical entity roles it is critical we associate as many chemicals as possible with a connection table (e.g. SMILES). If compound numbers are used without definition in a reaction (Fig. 3) then we cannot resolve the structures. The reaction extraction software is designed to run on an entire document, under those conditions these are less problematic as it is possible to resolve compound references to structures defined elsewhere in the document.

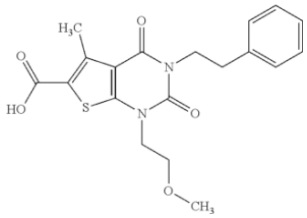
Step 2

[0944] To Compound 154 (30 mg, 0.057 mmol) and morpholine (0.015 mL, 0.171 mmol) in dichloromethane (0.9 mL), methanol (0.9 mL) and acetic acid (0.09 mL) solution was added picoline-borane complex (9.2 mg, 0.086 mmol), the mixture was stirred at room temperature for 1 hour. Saturated aqueous sodium hydrogen carbonate solution (20 mL) was added thereto, and the mixture was extracted with ethyl acetate (30 mL). The organic layer was washed with saturated brine (20 mL), and dried over anhydrous magnesium sulfate. After concentration, the residue was purified by amino silica gel chromatography (hexane-ethyl acetate) to obtain Compound 155 (23.2 mg, 68% yield) as brown foam.

[0945] LC/MS measurement conditions: (1), LC/MS (ESI): m/z=597.8 [M+H]+.

Example 15A

1-(2-Methoxyethyl)-5-methyl-2,4-dioxo-3-(2-phenylethyl)-1,2,3,4-tetrahydrothieno[2,3-d]pyrimidine-6-carboxylic acid

[0721] 

[0722] 75 ml of trifluoroacetic acid were added to a solution of 5.0 g (11.2 mmol) of the compound from Ex. 10A in 225 ml of dichloromethane, and the mixture was stirred at RT for 2 h. The reaction mixture was then concentrated to dryness on a rotary evaporator. The remaining residue was stirred in diethyl ether and filtered off with suction, and the solid was dried under high vacuum. 4.1 g (92% of theory) of the title compound were obtained.

Fig. 3. Reaction paragraphs that reference "Compound 154 and 155" (US20180162876A1 0944, train:0022) and "compound from Ex. 10A" (US10428083B2 0721 train:0026).

The ChEMU task provided paragraphs in isolation and so we cannot resolve the structure for referenced compounds. However by providing an additional context mapping of input paragraphs to the source patent document we can resolve more structures and improve our entity type assignment.

This mapping was obtained by finding a patent, within which the same text as the input paragraph was found. We used this mapping to call out to a LeadMine based web-service that automatically extracts the compound-id relationships from the queried document and returns these as a TSV file. The annotator can then use this additional context to assign structures.

3.4 Annotation Difference Viewer

Since we were adapting an existing tool to the annotation task we started with a baseline and then improved it incrementally as we identified differences in our output to what was expected by the gold standard annotations. All changes were carried out in the evaluation period of the task and a tool that could provide intuitive feedback quickly was required. Internally at NextMove a "reaction diff-viewer" web application is used to identify output changes between commits in our version control system and check for regressions. This web application was adapted to the ChEMU tasks and allowed us to rapidly identify differences from the gold standard and assess improvements and regressions between runs of the tool (Fig. 4). Over the course of the evaluation period the NER Exact F_1-Score on the training split was improved from an initial score 0.7933 to 0.9060.

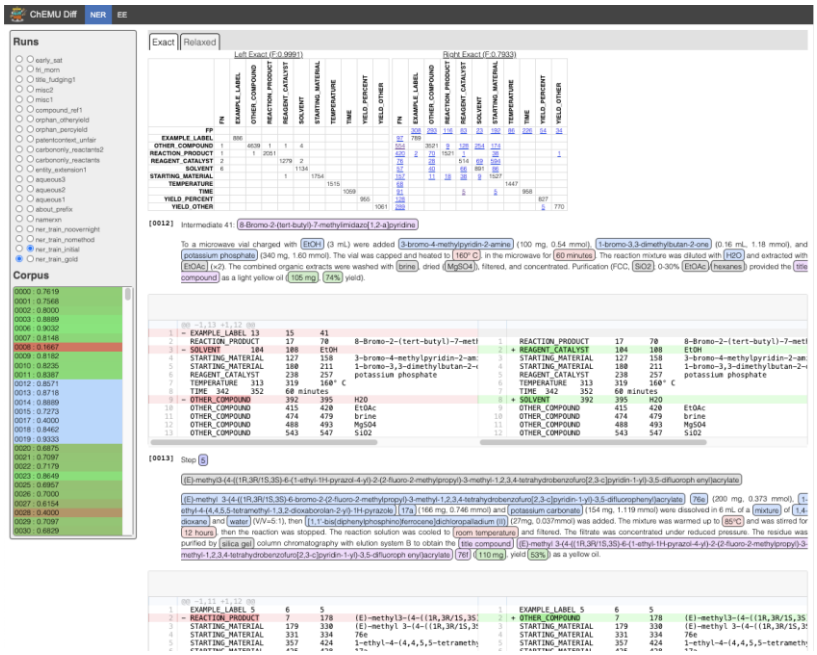


Fig. 4. Screenshot of the ChEMU Diff Viewer Web Application used to assist in matching gold standard annotations.

4 Results

Table 2. Results for Task 1: Named Entity Recognition

Run	Exact matching			Relaxed matching		
	F_1-score	Precision	Recall	F_1-score	Precision	Recall
With Patent Context	0.8983	0.9042	0.8924	0.9240	0.9301	0.9181
Without Patent Context	0.8977	0.9037	0.8918	0.9236	0.9294	0.9178

Table 3. Results for Task 2: Event extraction

F_1-score	Precision	Recall
0.8977	0.9441	0.8556

Table 4. Results for end-to-end (event extraction using named entities from Task 1)

Run	Exact matching			Relaxed matching		
	F_1-score	Precision	Recall	F_1-score	Precision	Recall
With Patent Context	0.8026	0.8492	0.7609	0.8196	0.8663	0.7777
Without Patent Context	0.8020	0.8486	0.7602	0.8188	0.8653	0.7771
With Patent Context + same algorithm as final task 2 submission	0.8255	0.8746	0.7816	0.8420	0.8909	0.7983

4.1 Complete Reactions

We evaluated the training split on subsets based on whether the tool thinks it understood the semantics of the paragraph and identified a reaction.

The first subset (Any Complete) includes any paragraphs where ≥1 complete reaction was found. A complete reaction passes various sanity checks (e.g. ≥2 reactant and ≥1 product structures) and must have been atom-mapped by NameRxn or Indigo. A refinement of this subset is paragraphs that only contain a single complete reaction (One Complete). Finally of those with a single complete reaction we evaluated only those that NameRxn recognized. We measured the last two categories with and without the patent context.

Table 5. NER scores for different subsets of the training split

Subset	Patent Context	Applicability	F_1-Score	Precision	Recall	Relaxed F_1-Score
Everything	Y	900 (100%)	0.90604	0.91602	0.89628	0.92517
Any Complete	Y	540 (60%)	0.92561	0.93429	0.91709	0.9397
One Complete	N	475 (52.7%)	0.92873	0.93675	0.92085	0.94080
One Complete	Y	529 (58.7%)	0.92621	0.93491	0.91768	0.94031
NameRxn	N	375 (41.6%)	0.93597	0.94431	0.92778	0.94717

Table 6. End-to-end scores for different subsets of the training split

Subset	Patent Context	Applicability	F_1-Score	Precision	Recall	Relaxed F_1-Score
Everything	Y	900 (100%)	0.82843	0.88384	0.77955	0.84443
Any Complete	Y	540 (60%)	0.84649	0.89556	0.80252	0.85909

One Complete	N	475 (52.7%)	0.85342	0.89961	0.81174	0.86486
One Complete	Y	529 (58.7%)	0.84877	0.89758	0.80499	0.86101
NameRxn	N	375 (41.6%)	0.86442	0.90775	0.82503	0.87583

4.2 Confusion Matrices

For the training and development splits we compared the expected entity type from the gold standard against the "predicted" entity type assigned by our tool (Tables 7 and 8).

Table 7. Confusion Matrix for Train Split (Relaxed)

		Predicted Entity Type										
		FN	EXAMPLE_LABEL	OTHER_COMPOUND	REACTION_PRODUCT	REAGENT_CATALYST	SOLVENT	STARTING_MATERIAL	TEMPERATURE	TIME	YIELD_PERCENT	YIELD_OTHER
Actual Entity Type	FP		8	59	21	23	6	35	23	31	4	35
	EXAMPLE_LABEL	45	840					1				
	OTHER_COMPOUND	228	2	4261	17	70	60	46				
	REACTION_PRODUCT	225	6	36	1759	2		65				1
	REAGENT_CATALYST	13		79	8	1098	19	68				
	SOLVENT	22		106		71	909	32				
	STARTING_MATERIAL	54		23	21	44	7	1632				
	TEMPERATURE	15							1501			
	TIME	6				6		4		1044		
	YIELD_PERCENT	25									930	
	YIELD_OTHER	58									3	1000

Table 8. Confusion Matrix for Dev Split (Relaxed)

Actual Entity Type \ Predicted Entity Type	FN	EXAMPLE_LABEL	OTHER_COMPOUND	REACTION_PRODUCT	REAGENT_CATALYST	SOLVENT	STARTING_MATERIAL	TEMPERATURE	TIME	YIELD_PERCENT	YIELD_OTHER
FP			14	8	5	1	6	5	8		13
EXAMPLE_LABEL	12	206									
OTHER_COMPOUND	67	1	985	5	12	9	9				
REACTION_PRODUCT	58	2	12	424			18				
REAGENT_CATALYST	1		19	2	248	7	12				
SOLVENT	4		16		15	209	6				
STARTING_MATERIAL	9		6	8	11		385				
TEMPERATURE	6							340			
TIME					1				251		
YIELD_PERCENT	4									224	
YIELD_OTHER	8									1	252

5 Discussion

As our solution is primarily based on an existing solution rather than being built or trained for this task, performance was primarily improved by adapting our existing output to match the annotation guidelines. This revealed a few notable quirks and inconsistencies. For some of these points, a specific example in the training split is referred to via its 4 digit paragraph number.

- “Overnight” isn’t considered to be a period of time in the gold standard
- The annotation guidelines indicated that inert gases should not be included in event annotations. However in the gold standard, more often than not these chemicals were included with no obvious distinction between the cases where they were and weren’t. As a result, despite having implemented detection for inert gases, in our final submission we made no distinction between these and other compounds.
- While the gold standard’s reaction events generally were similar in scope to ChemicalTagger’s, some events were very rarely annotated in the gold standard. A common example was a concentrate action e.g. “the filtrate was **concentrated** under vacuum”. We adjusted for this by not considering VB-CONCENTRATE to be a workup action trigger word.

- The annotation guidelines specified that if multiple temperatures were given for a reaction that only the lowest and highest should be retained. Due to the anticipated small difference in performance and high likelihood of important reaction condition information being excluded, this was not implemented.
- Presenting reaction paragraphs without the context of their originating patent can significantly complicate assigning roles as when a starting material was defined in a preceding experimental section, you will not know the structure of it, while from the complete patent this may have been possible. This means an atom mapping for the reaction likely won't be possible and hence the assignment of which chemicals are catalysts is complicated. Our "with patent context" runs were a proof of concept to investigate overcoming this limitation.
- The annotation guidelines do not distinguish between definitions of the product and label(s) associated with the product. This means that it's not uncommon for a single product reaction to have 3 product entities: a mention in the heading, a mention of the reaction outcome and a label associated with the product. This mismatches with our typical goal where the reaction data structure should only contain more than one product if a reaction yielded multiple compounds.
- 1034: "ice-water bath" occurs twice, water is tagged as "OTHER_COMPOUND" in only one. We would recognize this as an apparatus/equipment.
- Six train+dev paragraphs had no starting material, 0174,1036,1050,1055,1376 (US10258045B2)
- 1111: "target pale brown solid" is two entities "target" and "solid" 1198 "target white solid" is one. We recognized both as one.
- The trailing dot is sometimes omitted from the bounds for "aq." and "r.t." abbreviations (0185/0206 and 0833/1444).
- Boron tribromide is tagged as a STARTING_MATERIAL (0081) but doesn't contain any carbons, the guidelines list contributing a carbon to the product as a requirement for this entity type.
- Temperature ranges were handled inconsistently by the gold standard, necessitating adjusting the entity bounds to remove qualifiers and selectively splitting ranges. Closer correspondence with the gold standard was obtained by selectively splitting these ranges. Ranges were split if they were connected by "to" or "and", and the lower bound had an explicit unit attached. Removed quantity qualifiers included "approximately", "below", and "about".

Using the patent context information had little impact on the annotation scores, F_1 0.8983 vs F_1 0.8977 on the test split. However the benefit of the patent context is emphasised on how many complete reactions we can extract (Table 5). In the "one complete" subset we can generate 529 "complete reactions" instead of 475. The precision and recall is higher if we only consider paragraphs that we can extract a complete reaction. Further investigation is needed to determine if the tool performs better when there is a completable reaction or whether we extract more complete reactions due to better annotation.

Annotating the paragraphs with their source patent number reveals a large overlap between the training, development and test. This splitting is unrealistic as there are more similarities between how reactions are described within a document than between documents. As patents only apply to particular jurisdictions, it is common for multiple patents with essentially the same content to be filed in different jurisdictions. These patent equivalents should also be considered when splitting the data to avoid training and testing on different documents that nonetheless have the same content.

The confusion matrices (Table 7a and 7b) show the majority of mistakes are related to the chemical type (role) assignment. Unfortunately a miss-typed entity counts as both a FP and FN so eliminating these cases is desirable. Additional heuristics and statistics could help with distinguishing the non-product roles. An additional entity type confusion is seen with chemical entities and TIME entities. These cases are terms like “1h” and “2h” which the tool labels as plausible reference identifiers “add 12.2 g of 1h” but are actually time intervals “stirred for 1h”. These entities are considered too short and ambiguous for the LeadMine physical quantity grammar to recognize in general text. However with the additional context of the surrounding tags it is possible to rectify this in ChemicalTagger.

6 Conclusions

We present here a grammar based approach to chemical reaction extraction, demonstrating that this approach can achieve competitive performance when compared to contemporary machine learning approaches. A complete system using this approach had already been used to extract millions of reactions from patents resulting in valuable data resources.

References

1. Reeker, L.H., Zamora, E.M., Blower, P.E.: Specialized information extraction: automatic chemical reaction coding from English descriptions. In: Proceedings of the first conference on Applied natural language processing. pp. 109–116. Association for Computational Linguistics (1983).
2. Zamora, E.M., Blower Jr, P.E.: Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases. J. Chem. Inf. Comput. Sci. 24, 176–181 (1984).
3. He, J., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., Afzal, Z., Zhai, Z., Fang, B., Yoshikawa, H., Albahem, A., Cavedon, L., Cohn, T., Baldwin, T., Verspoor, K.: Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., and Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science (2020).

4. Lowe, D.M.: Extraction of chemical structures and reactions from the literature, http://www.repository.cam.ac.uk/handle/1810/244727, (2012).
5. Lowe, D.M.: Patent Reaction Extraction Project, https://github.com/dan2097/patent-reaction-extraction, last accessed 2020/07/09.
6. Lowe, D.M., Sayle, R.A.: LeadMine: A grammar and dictionary driven approach to entity recognition. Journal of Cheminformatics. 7, S5 (2015).
7. Lowe, D.M.: Chemical reactions from US patents (1976-Sep2016), https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, (2017). https://doi.org/10.6084/m9.figshare.5104873.v1.
8. NextMove Software: Pistachio, https://www.nextmovesoftware.com/pistachio, last accessed 2020/07/17.
9. Schneider, N., Lowe, D.M., Sayle, R.A., Tarselli, M.A., Landrum, G.A.: Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. J. Med. Chem. 59, 4385–4402 (2016). https://doi.org/10.1021/acs.jmedchem.6b00153.
10. Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P., Pande, V.: Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. ACS Cent. Sci. 3, 1103–1113 (2017). https://doi.org/10.1021/acscentsci.7b00303.
11. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C.A., Bekas, C., Lee, A.A.: Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. ACS Cent. Sci. 5, 1572–1583 (2019). https://doi.org/10.1021/acscentsci.9b00576.
12. Hawizy, L., Jessop, D.M., Adams, N., Murray-Rust, P.: ChemicalTagger: A tool for semantic text-mining in chemistry. J Cheminf. 3, 17 (2011). https://doi.org/10.1186/1758-2946-3-17.
13. Jessop, D.M., Adams, S., Willighagen, E.L., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. J Cheminf. 41 (2011). https://doi.org/10.1186/1758-2946-3-41.
14. Parr, T.: The definitive ANTLR reference: building domain-specific languages. Pragmatic Bookshelf (2007).
15. NextMove Software: NameRxn, https://www.nextmovesoftware.com/namerxn.html, last accessed 2020/11/07.
16. EPAM Systems: Indigo Toolkit, https://lifescience.opensource.epam.com/indigo/, last accessed 2020/07/11.
17. Clark, A.M., McEwen, L.R., Gedeck, P., Bunin, B.A.: Capturing mixture composition: an open machine-readable format for representing mixed substances. Journal of Cheminformatics. 11, 33 (2019). https://doi.org/10.1186/s13321-019-0357-4.
18. Gushurst, A.J., Nourse, J.G., Hounshell, W.D., Leland, B.A., Raich, D.G.: The substance module: the representation, storage, and searching of complex structures. J. Chem. Inf. Comput. Sci. 31, 447–454 (1991). https://doi.org/10.1021/ci00004a003.