NLPatVCU CLEF 2020: ChEMU Shared Task System Description

Darshini Mahendran, Gabrielle Gurdin, Nastassja Lewinski, Christina Tang &

Bridget T. McInnes

Virginia Commonwealth University

Outline

1. Introduction
2. Data
3. Methods
4. Results and Error Analysis
5. Conclusion and Future Work

Introduction

VCU College of Engineering

ChEMU 2020: Cheminformatics Elsevier Melbourne University

- Shared task for Information Extraction from Chemical Patents

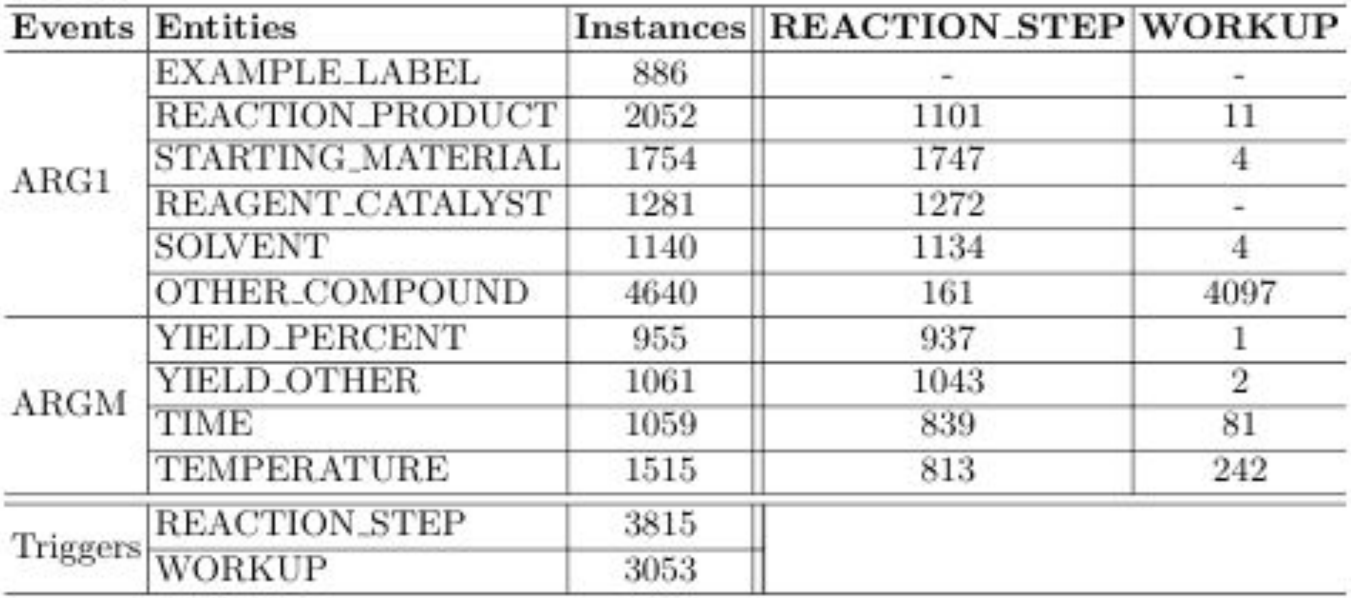
- ChEMU proposes two key information extraction tasks over chemical reactions from patent documents

- Tasks:
  - ***Task 1:*** Named Entity Recognition (NER) involves identifying chemical compounds as well as their types in context, i.e., to assign the label of a chemical compound according to the role which the compound plays within a chemical reaction
  - ***Task 2:*** Event Extraction (EE) over chemical reactions involves event trigger detection and argument recognition.

Data

VCU
College of Engineering

Data

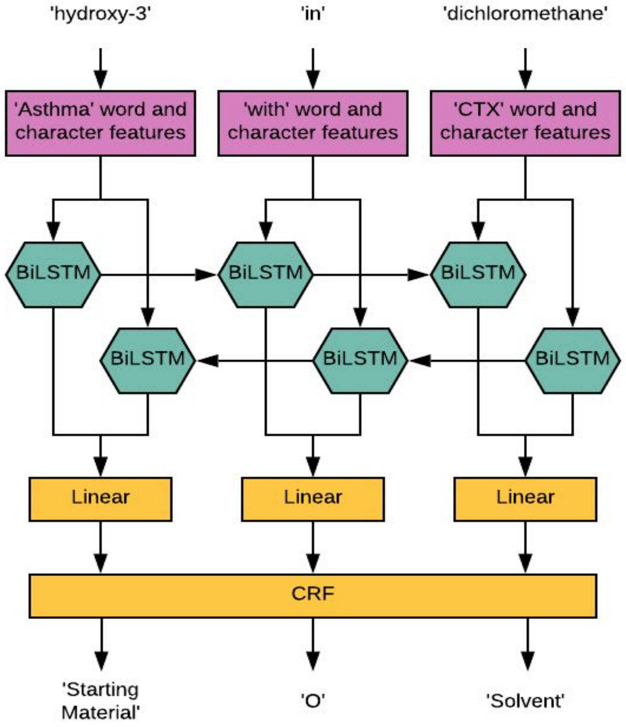


Events	Entities	Instances	REACTION_STEP	WORKUP
ARG1	EXAMPLE_LABEL	886	-	-
	REACTION_PRODUCT	2052	1101	11
	STARTING_MATERIAL	1754	1747	4
	REAGENT_CATALYST	1281	1272	-
	SOLVENT	1140	1134	4
	OTHER_COMPOUND	4640	161	4097
ARGM	YIELD_PERCENT	955	937	1
	YIELD_OTHER	1061	1043	2
	TIME	1059	839	81
	TEMPERATURE	1515	813	242
Triggers	REACTION_STEP	3815		
	WORKUP	3053		

Method: Named Entity Recognition

VCU
College of Engineering

NER & Trigger Detection: Algorithm

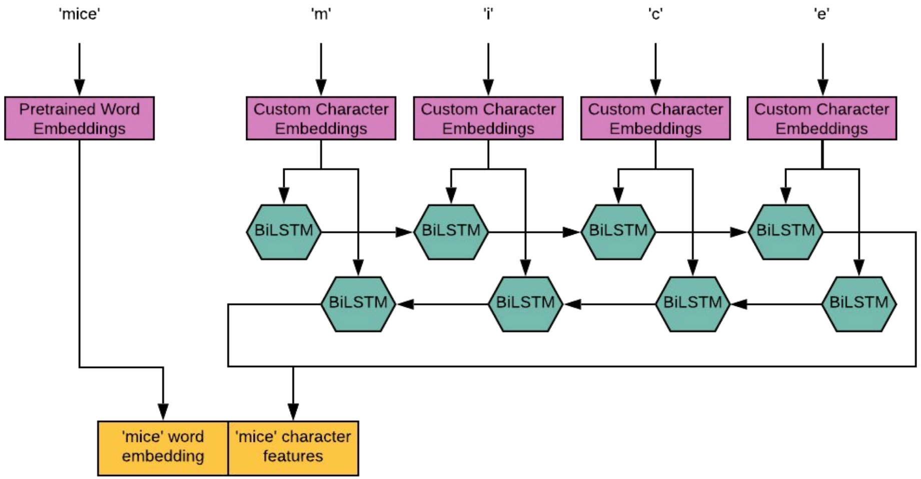


- Bidirectional Long Short Term Memory (Bi-LSTM) units with a Conditional Random Field (CRF) output layer

- BiLSTMs - type of Recurrent Neural Network
  - 2 sources of input: their current state and their past states

- A linear-chain CRF is used to assign the final class probability

NER & Trigger Detection: Feature Representation


'mice'
'm'
'i'
'c'
'e'
Pretrained Word Embeddings
Custom Character Embeddings
Custom Character Embeddings
Custom Character Embeddings
Custom Character Embeddings
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
'mice' word embedding
'mice' character features

Input to our model is pre-trained word embeddings in combination with character embeddings

- Word2vec embeddings
  - ChemPatent: Trained over a collection of 84,076 full patent documents
  - WikiPubMed: Trained over Wikipedia and PubMed articles

- Character embeddings - learned using a biLSTM and concatenated into the word2vec embedding

Method:
Event Extraction

VCU

College of Engineering

Event Extraction

- To identify trigger words - NER system discussed previously

- To identify the chemical arguments between the trigger words and the entities
  - Rule-based Method
  - Convolutional Neural Network (CNN)-based Method

Rule-based method

- Utilizes co-location information of trigger words to determine with respect to entity if the word is referring to trigger word or not
    - Breadth-first search (BFS) algorithm is used here for traversal
    - For each entity, both sides are traversed until the closest occurrence of the trigger word is found using the provided span values of the entities

Rule-based

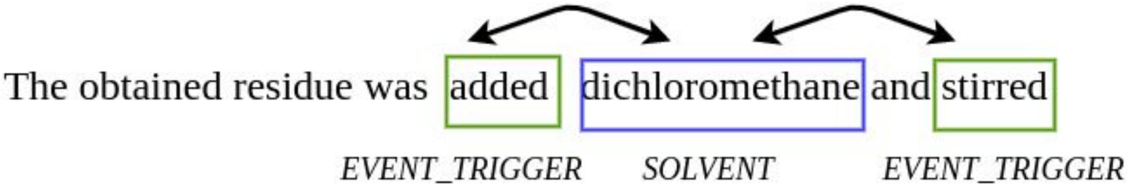
Different traversal techniques are applied and best traversal technique for each relation type is determined

- traverse left side only
- traverse right side only
- traverse left first then right
- traverse right first then left
- traverse both sides within a sentence

Rule-based

Different traversal techniques are applied and best traversal technique for each relation type is determined

- [✓ traverse left side only]
- traverse right side only
- traverse left first then right
- traverse right first then left
- traverse both sides within a sentence

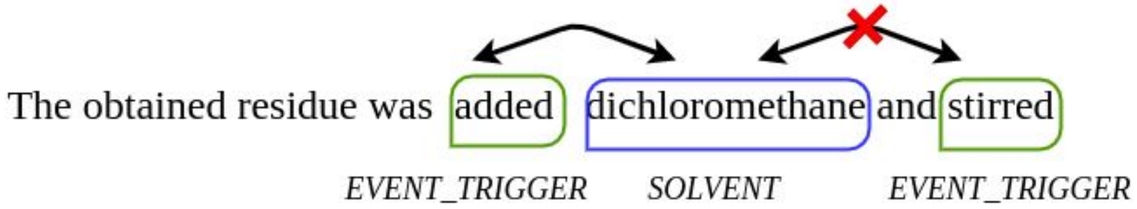


The obtained residue was added dichloromethane and stirred

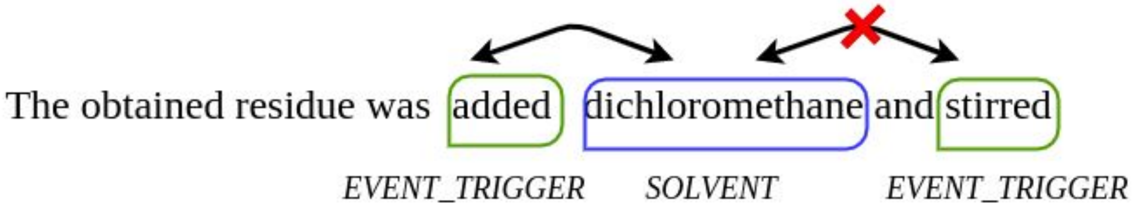
Rule-based

Different traversal techniques are applied and best traversal technique for each relation type is determined

- ○ [✓ traverse left side only]
- ○ traverse right side only
- ○ traverse left first then right
- ○ traverse right first then left
- ○ traverse both sides within a sentence



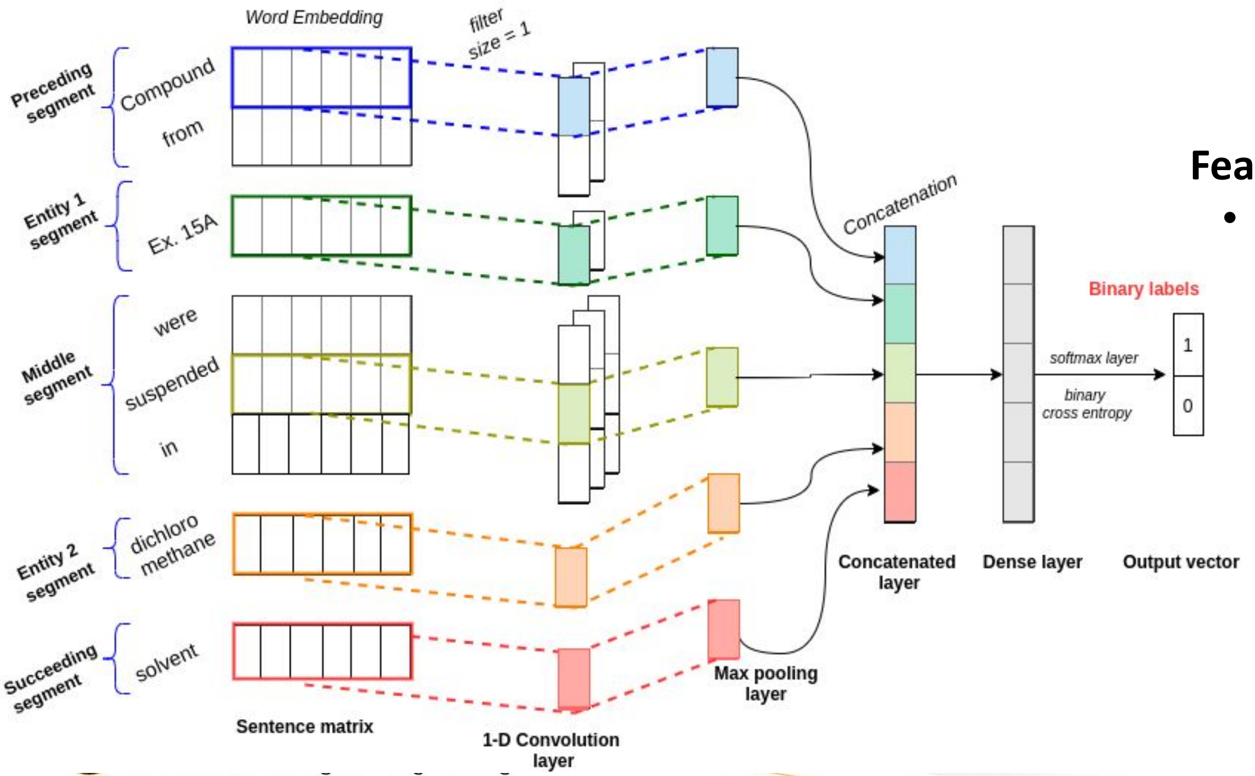
The obtained residue was added dichloromethane and stirred



The obtained residue was added dichloromethane and stirred

EVENT_TRIGGER *SOLVENT* *EVENT_TRIGGER*

CNN-based



Algorithm:

- for each *Trigger word-Entity pair* we perform a binary classification

Feature representation:

- ChemPatent - Trained over 84,076 patents

VCU NLP LAB

Results & Analysis

VCU

College of Engineering

Evaluation Metrics

- **Precision:** ratio between correctly predicted mentions over the total set of predicted mentions for a specific entity

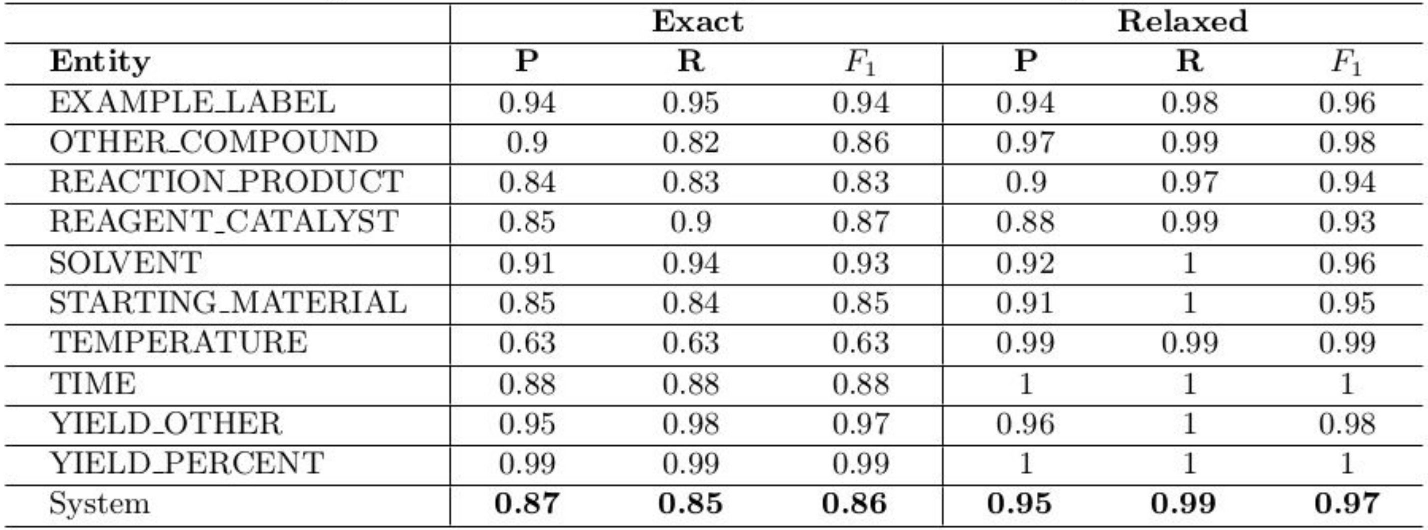
- **Recall:** ratio of correctly predicted mentions over the actual number of mentions

- **F-1 score:** harmonic mean between precision and recall

- For Task 1, we report both the exact and relaxed results for each entity category
  - *exact evaluation:* two annotations are equal only if they have the same tag with exactly matching spans
  - *relaxed evaluation:* two annotations are equal if they share the same tag and their spans overlap with each other.

Task 1: NER Results (Run 1)

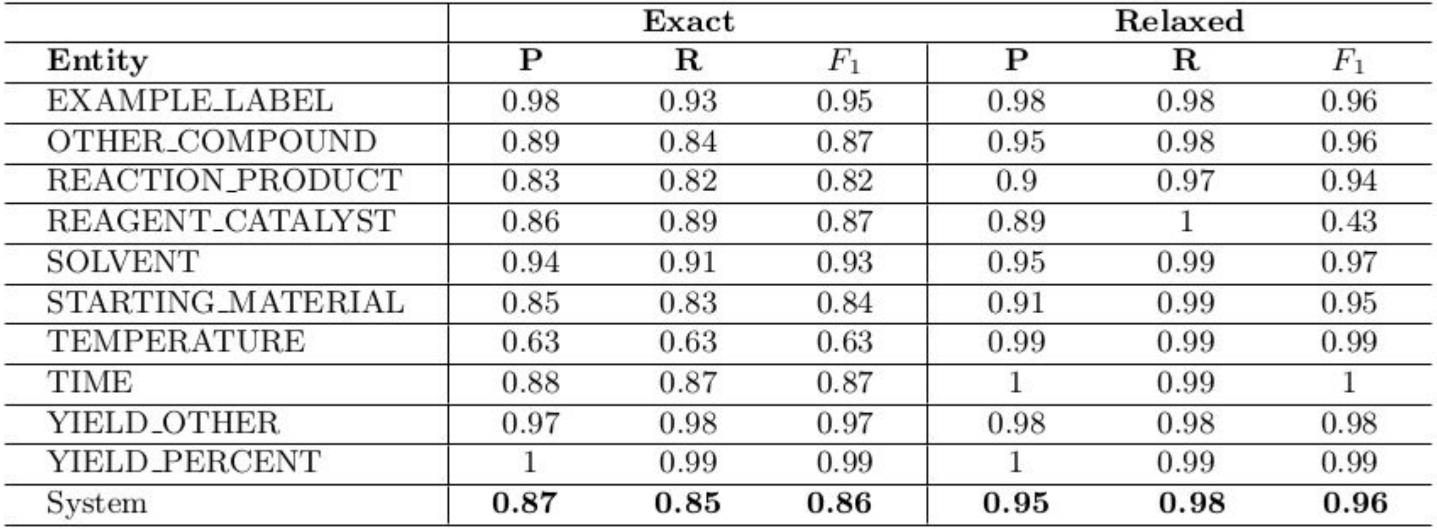
Run 1 - Model trained over the training data using the biLSTM+CRF with the CheMU Patent embeddings



	Exact			**Relaxed**		
Entity	**P**	**R**	F_1	**P**	**R**	F_1
EXAMPLE_LABEL	0.94	0.95	0.94	0.94	0.98	0.96
OTHER_COMPOUND	0.9	0.82	0.86	0.97	0.99	0.98
REACTION_PRODUCT	0.84	0.83	0.83	0.9	0.97	0.94
REAGENT_CATALYST	0.85	0.9	0.87	0.88	0.99	0.93
SOLVENT	0.91	0.94	0.93	0.92	1	0.96
STARTING_MATERIAL	0.85	0.84	0.85	0.91	1	0.95
TEMPERATURE	0.63	0.63	0.63	0.99	0.99	0.99
TIME	0.88	0.88	0.88	1	1	1
YIELD_OTHER	0.95	0.98	0.97	0.96	1	0.98
YIELD_PERCENT	0.99	0.99	0.99	1	1	1
System	**0.87**	**0.85**	**0.86**	**0.95**	**0.99**	**0.97**

Task 1: NER Results (Run 2)

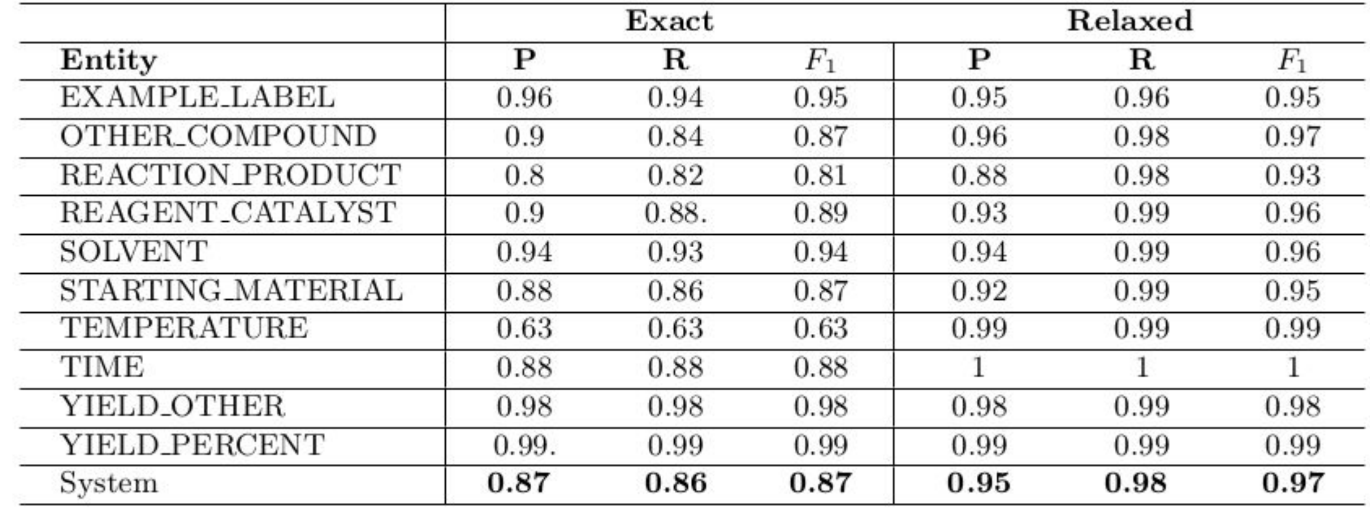
Run 2 - Model trained over the training data using the biLSTM+CRF with the WikiPubmed embeddings



	Exact			Relaxed		
Entity	**P**	**R**	F_1	**P**	**R**	F_1
EXAMPLE_LABEL	0.98	0.93	0.95	0.98	0.98	0.96
OTHER_COMPOUND	0.89	0.84	0.87	0.95	0.98	0.96
REACTION_PRODUCT	0.83	0.82	0.82	0.9	0.97	0.94
REAGENT_CATALYST	0.86	0.89	0.87	0.89	1	0.43
SOLVENT	0.94	0.91	0.93	0.95	0.99	0.97
STARTING_MATERIAL	0.85	0.83	0.84	0.91	0.99	0.95
TEMPERATURE	0.63	0.63	0.63	0.99	0.99	0.99
TIME	0.88	0.87	0.87	1	0.99	1
YIELD_OTHER	0.97	0.98	0.97	0.98	0.98	0.98
YIELD_PERCENT	1	0.99	0.99	1	0.99	0.99
System	**0.87**	**0.85**	**0.86**	**0.95**	**0.98**	**0.96**

Task 1: NER Results (Run 3)

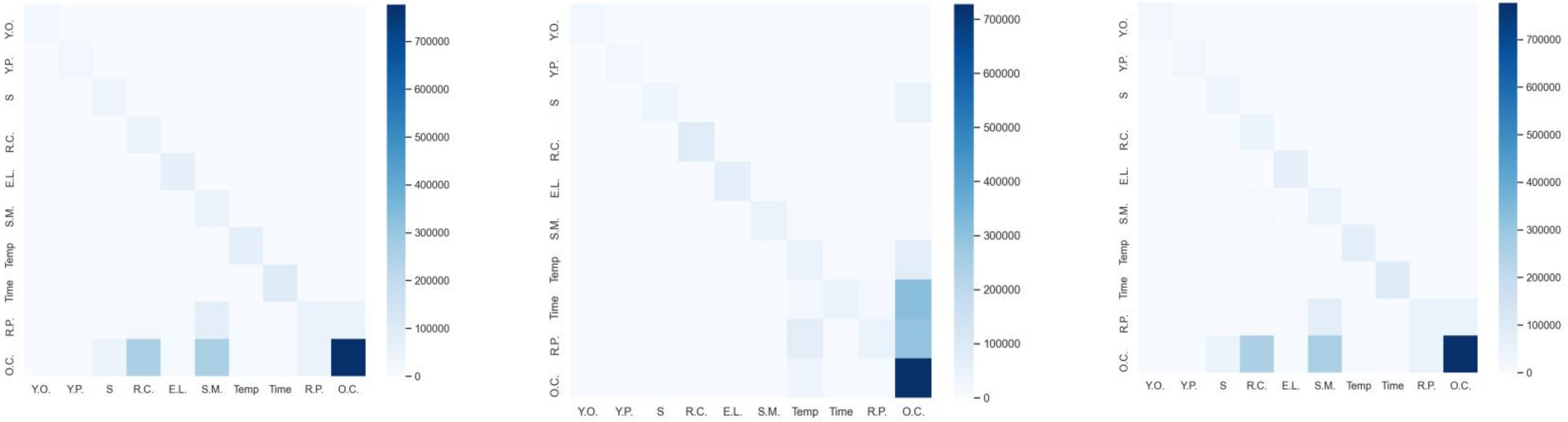
Run 3 - model trained over the training and development data combined with the biLSTM+CRF using the WikiPubmed embeddings.

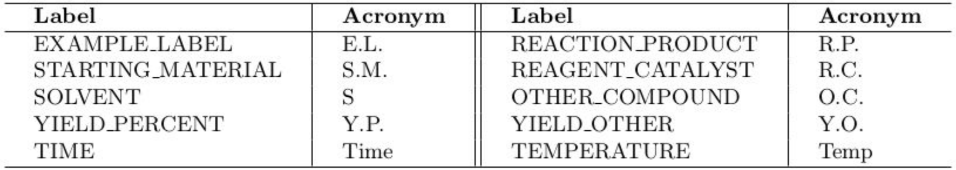


	Exact			Relaxed		
Entity	**P**	**R**	F_1	**P**	**R**	F_1
EXAMPLE_LABEL	0.96	0.94	0.95	0.95	0.96	0.95
OTHER_COMPOUND	0.9	0.84	0.87	0.96	0.98	0.97
REACTION_PRODUCT	0.8	0.82	0.81	0.88	0.98	0.93
REAGENT_CATALYST	0.9	0.88.	0.89	0.93	0.99	0.96
SOLVENT	0.94	0.93	0.94	0.94	0.99	0.96
STARTING_MATERIAL	0.88	0.86	0.87	0.92	0.99	0.95
TEMPERATURE	0.63	0.63	0.63	0.99	0.99	0.99
TIME	0.88	0.88	0.88	1	1	1
YIELD_OTHER	0.98	0.98	0.98	0.98	0.99	0.98
YIELD_PERCENT	0.99.	0.99	0.99	0.99	0.99	0.99
System	**0.87**	**0.86**	**0.87**	**0.95**	**0.98**	**0.97**

Task 1: Error Analysis

Confusion matrices for all 3 runs over the testing dataset (rows : annotated entities , columns: predicted entities)

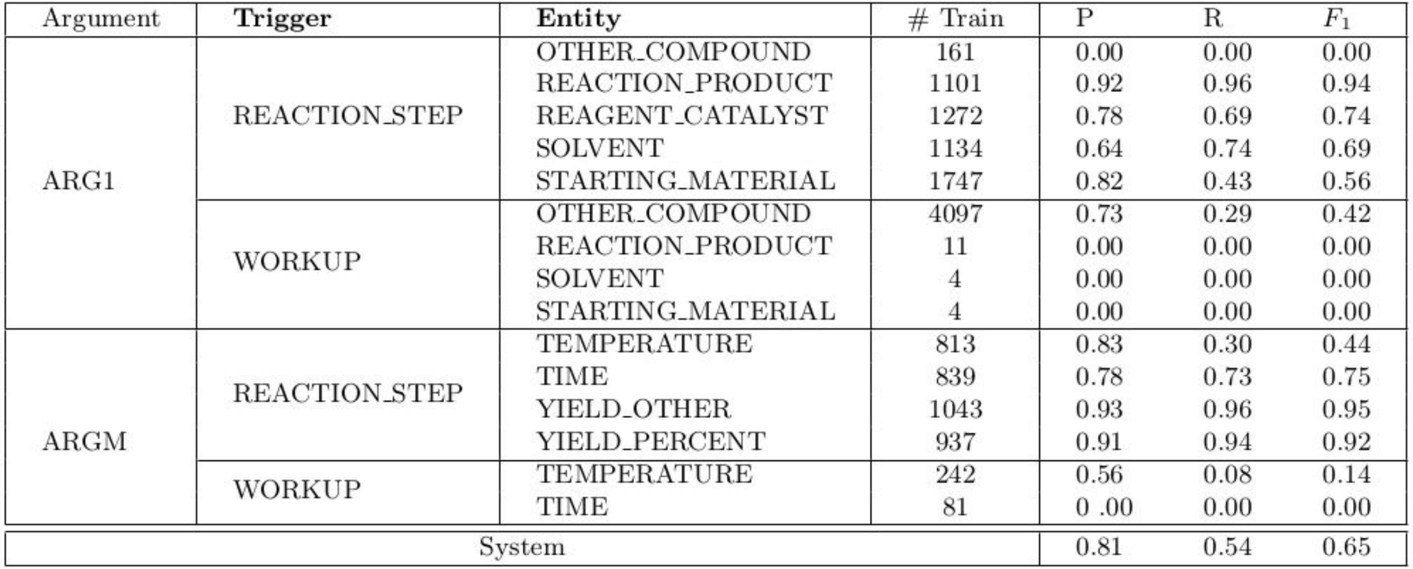




Label	Acronym	Label	Acronym
EXAMPLE_LABEL	E.L.	REACTION_PRODUCT	R.P.
STARTING_MATERIAL	S.M.	REAGENT_CATALYST	R.C.
SOLVENT	S	OTHER_COMPOUND	O.C.
YIELD_PERCENT	Y.P.	YIELD_OTHER	Y.O.
TIME	Time	TEMPERATURE	Temp

Task 2: Event extraction (Run 1)

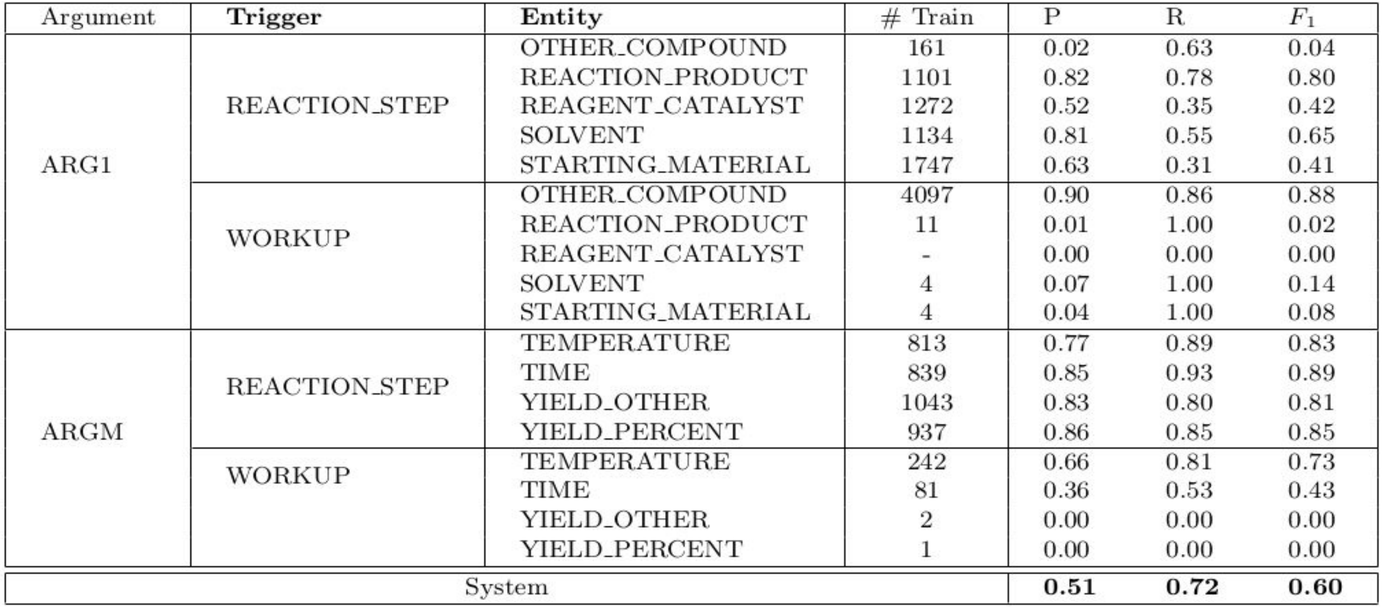
Run 1 - CNN-based system with trigger words identified using NER system trained with CheMU patent embeddings



Argument	Trigger	Entity	# Train	P	R	F_1
ARG1	REACTION_STEP	OTHER_COMPOUND	161	0.00	0.00	0.00
		REACTION_PRODUCT	1101	0.92	0.96	0.94
		REAGENT_CATALYST	1272	0.78	0.69	0.74
		SOLVENT	1134	0.64	0.74	0.69
		STARTING_MATERIAL	1747	0.82	0.43	0.56
	WORKUP	OTHER_COMPOUND	4097	0.73	0.29	0.42
		REACTION_PRODUCT	11	0.00	0.00	0.00
		SOLVENT	4	0.00	0.00	0.00
		STARTING_MATERIAL	4	0.00	0.00	0.00
ARGM	REACTION_STEP	TEMPERATURE	813	0.83	0.30	0.44
		TIME	839	0.78	0.73	0.75
		YIELD_OTHER	1043	0.93	0.96	0.95
		YIELD_PERCENT	937	0.91	0.94	0.92
	WORKUP	TEMPERATURE	242	0.56	0.08	0.14
		TIME	81	0 .00	0.00	0.00
System				0.81	0.54	0.65

Task 2: Event extraction (Run 2)

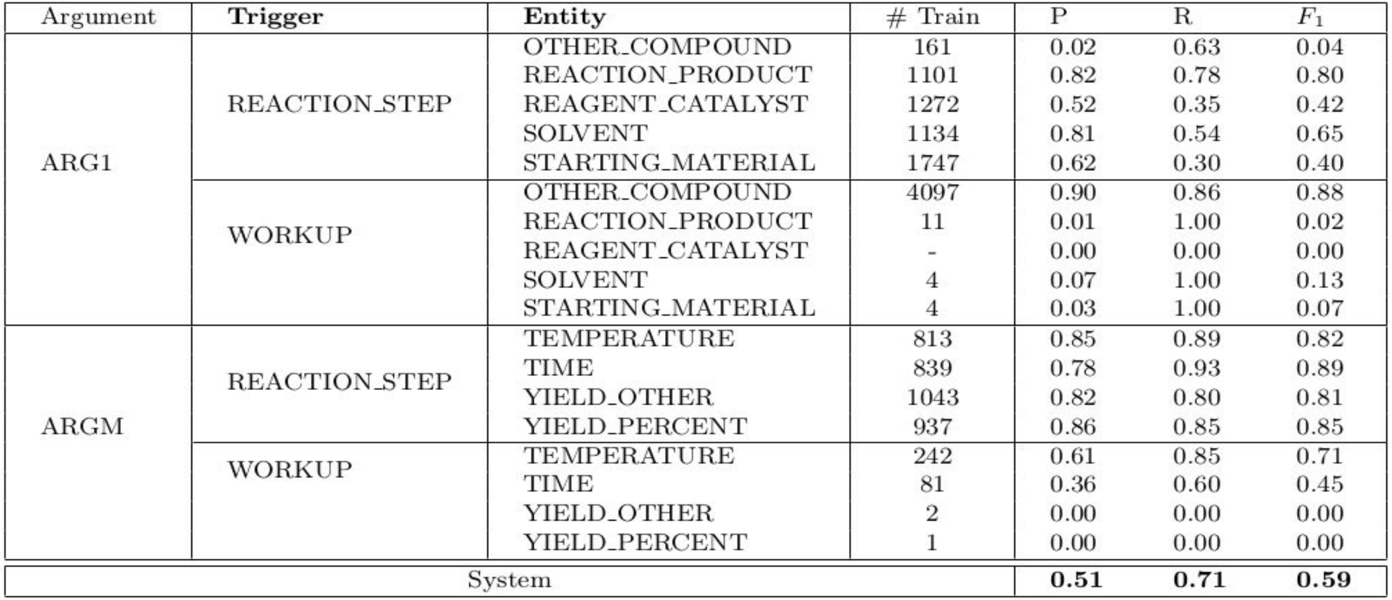
Run 2 - Rule-based system with trigger words identified using NER system trained with CheMU patent embeddings



Argument	Trigger	Entity	# Train	P	R	F_1
ARG1	REACTION_STEP	OTHER_COMPOUND	161	0.02	0.63	0.04
		REACTION_PRODUCT	1101	0.82	0.78	0.80
		REAGENT_CATALYST	1272	0.52	0.35	0.42
		SOLVENT	1134	0.81	0.55	0.65
		STARTING_MATERIAL	1747	0.63	0.31	0.41
	WORKUP	OTHER_COMPOUND	4097	0.90	0.86	0.88
		REACTION_PRODUCT	11	0.01	1.00	0.02
		REAGENT_CATALYST	-	0.00	0.00	0.00
		SOLVENT	4	0.07	1.00	0.14
		STARTING_MATERIAL	4	0.04	1.00	0.08
ARGM	REACTION_STEP	TEMPERATURE	813	0.77	0.89	0.83
		TIME	839	0.85	0.93	0.89
		YIELD_OTHER	1043	0.83	0.80	0.81
		YIELD_PERCENT	937	0.86	0.85	0.85
	WORKUP	TEMPERATURE	242	0.66	0.81	0.73
		TIME	81	0.36	0.53	0.43
		YIELD_OTHER	2	0.00	0.00	0.00
		YIELD_PERCENT	1	0.00	0.00	0.00
System				**0.51**	**0.72**	**0.60**

Task 2: Event extraction (Run 3)

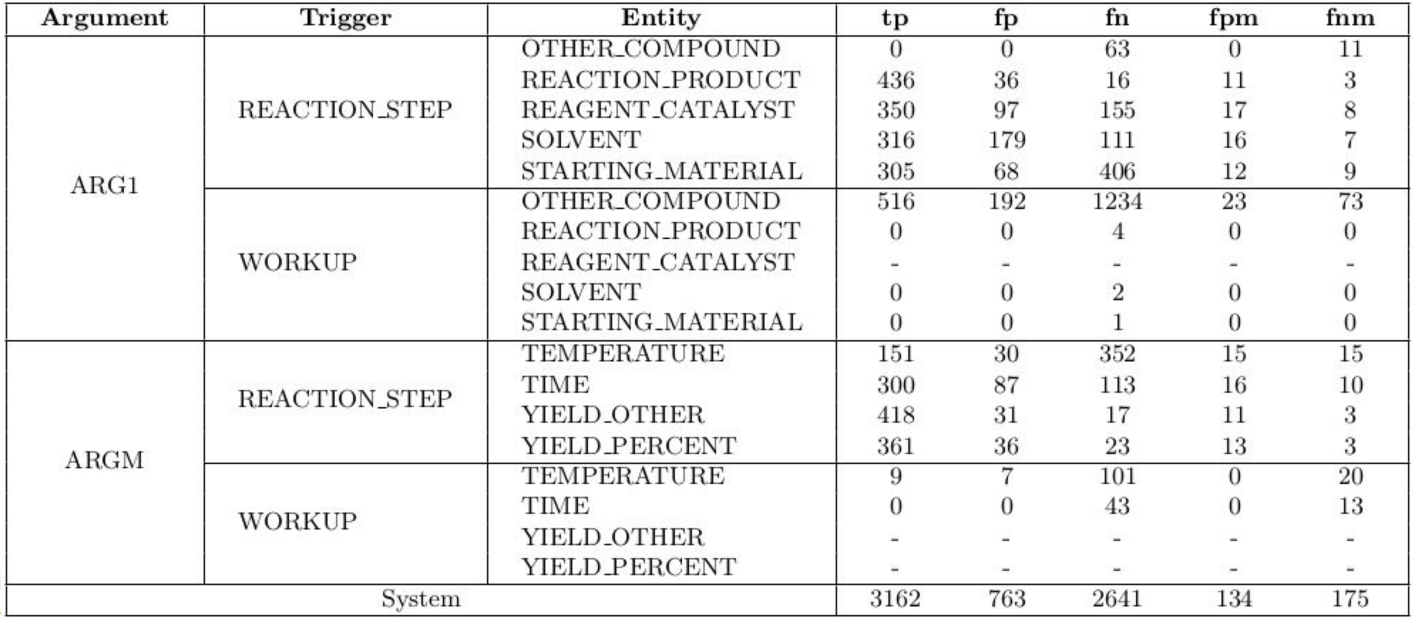
Run 3 - Rule-based system with trigger words identified using NER system trained with WikiPubmed embeddings



Argument	**Trigger**	**Entity**	# Train	P	R	F_1
ARG1	REACTION_STEP	OTHER_COMPOUND	161	0.02	0.63	0.04
		REACTION_PRODUCT	1101	0.82	0.78	0.80
		REAGENT_CATALYST	1272	0.52	0.35	0.42
		SOLVENT	1134	0.81	0.54	0.65
		STARTING_MATERIAL	1747	0.62	0.30	0.40
	WORKUP	OTHER_COMPOUND	4097	0.90	0.86	0.88
		REACTION_PRODUCT	11	0.01	1.00	0.02
		REAGENT_CATALYST	-	0.00	0.00	0.00
		SOLVENT	4	0.07	1.00	0.13
		STARTING_MATERIAL	4	0.03	1.00	0.07
ARGM	REACTION_STEP	TEMPERATURE	813	0.85	0.89	0.82
		TIME	839	0.78	0.93	0.89
		YIELD_OTHER	1043	0.82	0.80	0.81
		YIELD_PERCENT	937	0.86	0.85	0.85
	WORKUP	TEMPERATURE	242	0.61	0.85	0.71
		TIME	81	0.36	0.60	0.45
		YIELD_OTHER	2	0.00	0.00	0.00
		YIELD_PERCENT	1	0.00	0.00	0.00
System				**0.51**	**0.71**	**0.59**

Task 2: Error Analysis

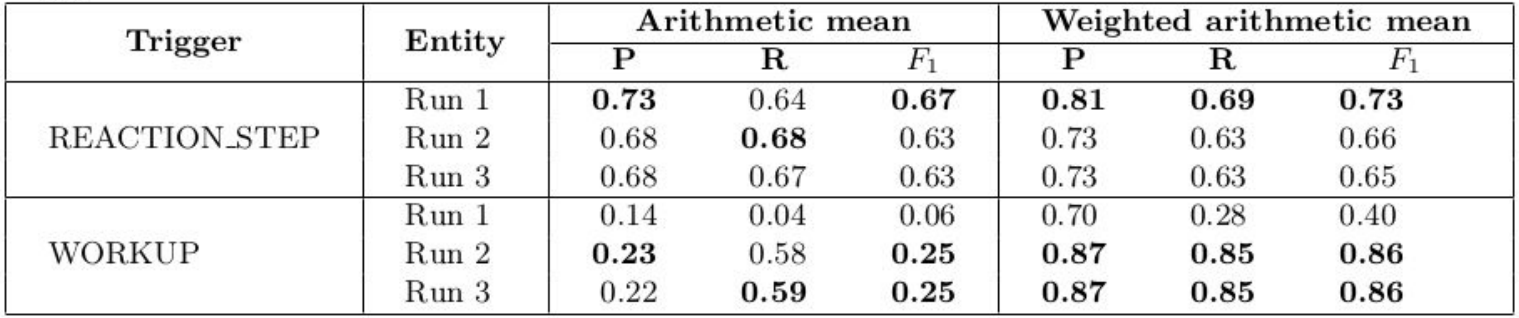
Error analysis for the CNN model trained with ChemPatent embedding



Argument	Trigger	Entity	tp	fp	fn	fpm	fnm
ARG1	REACTION_STEP	OTHER_COMPOUND	0	0	63	0	11
		REACTION_PRODUCT	436	36	16	11	3
		REAGENT_CATALYST	350	97	155	17	8
		SOLVENT	316	179	111	16	7
		STARTING_MATERIAL	305	68	406	12	9
	WORKUP	OTHER_COMPOUND	516	192	1234	23	73
		REACTION_PRODUCT	0	0	4	0	0
		REAGENT_CATALYST	-	-	-	-	-
		SOLVENT	0	0	2	0	0
		STARTING_MATERIAL	0	0	1	0	0
ARGM	REACTION_STEP	TEMPERATURE	151	30	352	15	15
		TIME	300	87	113	16	10
		YIELD_OTHER	418	31	17	11	3
		YIELD_PERCENT	361	36	23	13	3
	WORKUP	TEMPERATURE	9	7	101	0	20
		TIME	0	0	43	0	13
		YIELD_OTHER	-	-	-	-	-
		YIELD_PERCENT	-	-	-	-	-
System			3162	763	2641	134	175

Task 2: Error Analysis

Arithmetic and Weighted arithmetic mean of the performance of the trigger words for each run



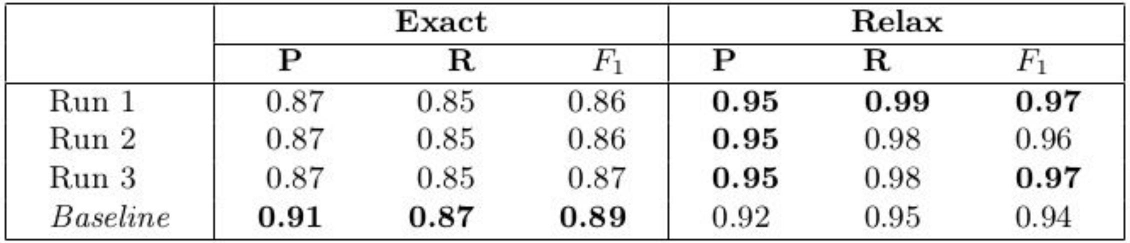
Trigger	Entity	Arithmetic mean			Weighted arithmetic mean		
		P	R	F_1	P	R	F_1
REACTION_STEP	Run 1	**0.73**	0.64	**0.67**	**0.81**	**0.69**	**0.73**
	Run 2	0.68	**0.68**	0.63	0.73	0.63	0.66
	Run 3	0.68	0.67	0.63	0.73	0.63	0.65
WORKUP	Run 1	0.14	0.04	0.06	0.70	0.28	0.40
	Run 2	**0.23**	0.58	**0.25**	**0.87**	**0.85**	**0.86**
	Run 3	0.22	**0.59**	**0.25**	**0.87**	**0.85**	**0.86**

VCU College of Engineering

VCU NLP LAB

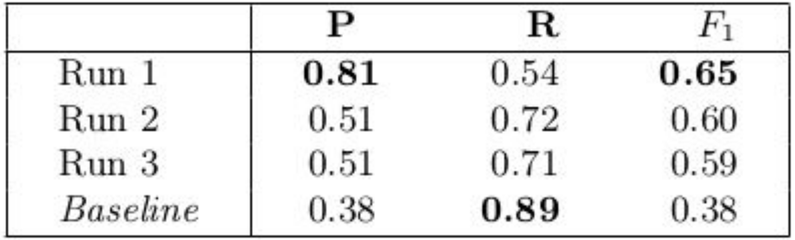
Comparison with the baseline

- *Task 1:*



	Exact			Relax		
	P	**R**	F_1	**P**	**R**	F_1
Run 1	0.87	0.85	0.86	**0.95**	**0.99**	**0.97**
Run 2	0.87	0.85	0.86	**0.95**	0.98	0.96
Run 3	0.87	0.85	0.87	**0.95**	0.98	**0.97**
Baseline	**0.91**	**0.87**	**0.89**	0.92	0.95	0.94

- *Task 2:*



	P	**R**	F_1
Run 1	**0.81**	0.54	**0.65**
Run 2	0.51	0.72	0.60
Run 3	0.51	0.71	0.59
Baseline	0.38	**0.89**	0.38

Conclusion
&
Future work

VCU College of Engineering

Task 1: Conclusions

- Evaluated 3 biLSTM+CRF models over different pre-trained word embeddings
    - models did not outperform the baseline model when evaluating exact span matches
    - models outperformed the baseline when evaluating in relaxed mode
- Errors primarily occurred because of issues with the model distinguishing between different entity labels
    - Example: mislabeling entities annotated as OTHER_COMPOUND for more specific labels, like REACTION_PRODUCT or STARTING_MATERIAL

Task 2: Conclusions

- Used a CNN-based model and 2 rule-based models to extract events
  - All 3 models outperformed the baseline model
  - CNN-based method outperforms the rule-based methods, especially with the REACTION_STEP classes as those classes have more instances to train on
  - Rule-based methods do not require training instances to train they perform better with WORKUP classes

Future Work

- Explore additional segment-CNN architectures
  - incorporate CRF layer while concatenating segments
  - incorporate biLSTM
  - incorporate transformer with attention mechanism

- Explore different feature representations :
  - *Feature-based representation*
    - incorporate semantic similarity, relatedness and association
  - *Featureless representation*
    - Character embeddings
    - Combine word and character embeddings
    - Contextual representation (e.g. BERT, ELMO)

Thank you!