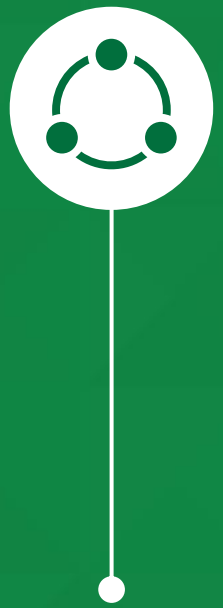
CPAIOR 2023
20th Intl. Conf. on the Integration of Constraint Programming Artificial Intelligence and Operations Research, Nice, France



Ner4Opt: Named Entity Recognition for Optimization Modelling from Natural Language

Parag Dakle[1], **Serdar Kadıoğlu[1,2]**, Karthik Uppuluri[1] Regina Politi[1], Preethi Raghavan[1], SaiKrishna Rallabandi[1], Ravisutha Srinivasamurthy[1]

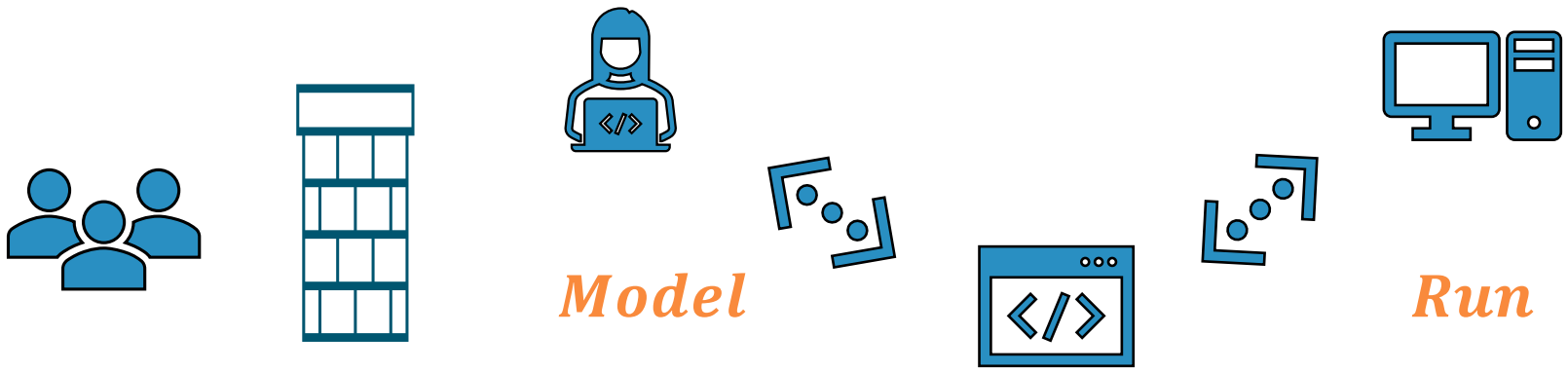
[1] AI Center of Excellence, Fidelity Investments
[2] Computer Science Department, Brown University

skadio.github.io

Introduction

Optimization Technology

- **Optimization** technology enjoys a wide range of applications
- Over the years, **dramatical speed-ups** enabled by theoretical and practical advances
- The overall **process** of modeling and solving problems remained the same for decades



Introduction

Ner4Opt: Named Entity Recognition for Optimization Modelling

- Envision **automated modeling assistant** to turn natural language into optimization formulations
- Necessary **building block**: finding key pieces of information relevant to optimization
- **Ner4Opt**: extracting optimization-related information such as the objective, constraints, and variables from free-form natural language text

Ner4Opt Problem

Library Demo

https://huggingface.co/spaces/skadio/Ner4Opt

Ner4Opt Problem Definition

Lexical and Semantic Solutions

Hybridization, Augmentation and Fine-Tuning

Extends our previous work
Dakle et. al. A Hybrid Model for Named Entity Recognition in Optimization Problems, NeurIPS'22

Ner4Opt: Named Entity Recognition for Optimization Modelling

Problem Definition and Optimization Entities

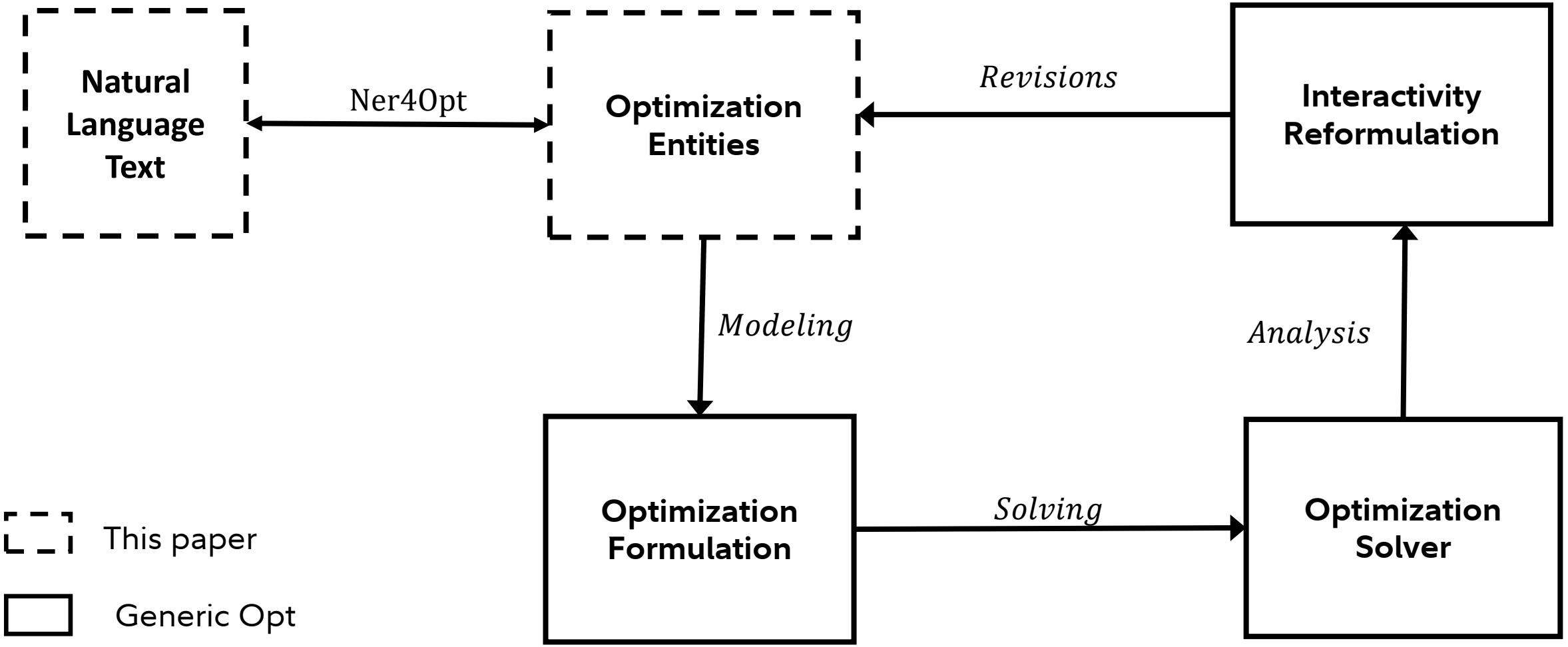
Given a sequence of tokens $s = \langle w_1, w_2, \ldots, w_n \rangle$, the goal of Ner4Opt is to output a list of tuples $\langle l_s, l_e, t \rangle$ each of which is a named entity specified in s. Here, $l_s \in [1, n]$ and $l_e \in [1, n]$ are the start and end indexes of a named entity while t is the entity type from a predefined category set of constructs related to optimization.

Predefined Optimization Entities

- $\boldsymbol{VAR}$: The variables of the problem – two advertising channels: **morning tv show** and **social media**
- $\boldsymbol{CONST\_DIR}$: The constraint direction – social media spots needs to be **at least** 30
- $\boldsymbol{LIMIT}$: Limits of constraints – plan at least **4** but no more than **7** morning show spots
- $\boldsymbol{OBJ\_NAME}$: The objective variable – maximize the **reach** of the campaign
- $\boldsymbol{OBJ\_DIR}$: The direction of optimization – **maximize** the reach of the campaign
- $\boldsymbol{PARAM}$: The parameters of the problem – costs the company $**1,000** to run advertisement spots

High-Level Architecture

Ner4Opt in the big picture



Ner vs. Ner4Opt

Challenges of Optimization Context

- ❑ NER for **information retrieval**, question answering, and machine translation
- ❑ **Multi-sentence word problem** with high-level of compositionality, ambiguity, variability
- ❑ Ner4Opt must be **domain agnostic** and generalize to new instances and applications
- ❑ **Extremely limited training data**. Even human annotation requires expertise.
  Must operate on low-resource regime

Chinchor et. al.: Message Understanding-7 named entity task definition, MUC, 1998

Solving the Ner4Opt

Classical and Modern NLP and their Hybridization

Conditional Random Field

Augmentation and Fine-Tuning

Solution Components

Features – Models – Data Centric Approach

1	**Feature Extraction, Engineering, and Learning**	Classical and semantic models to extract features for tokens while leveraging optimization context
2	**Conditional Random Field Neural Networks**	Linear chain conditional random field or fully connected network as the modeling component
3	**Data Augmentation Fine Tuning LLMs**	Augment the data set and fine-tune pre-trained large-language models

Conditional Random Field

Brief Introduction

Given an input sequence of tokens $\mathbf{x_i}$ and a set of feature extraction functions $\mathbf{f_j}$ at each token position, a **conditional random field** models a conditional probability distribution of labels $\mathbf{y_i}$ that can be assigned to appropriate segments in x.

$$D = [(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_d, y_d)] \quad i.i.d\ training\ examples \tag{1}$$

CRF

$$score\,(y|x) = \sum_{j=1}^{m}\sum_{i=1}^{n} w_j f_j\,(x, i, y_i, y_{i-1}) \tag{2}$$

$$p(y|x) = \frac{\exp^{score(y|x)}}{\sum_{y'} \exp^{score(y'|x)}} \tag{3}$$

$$L(w, D) = -\sum_{k=1}^{d} log\left[p(y^k|x^k)\right] \tag{4}$$

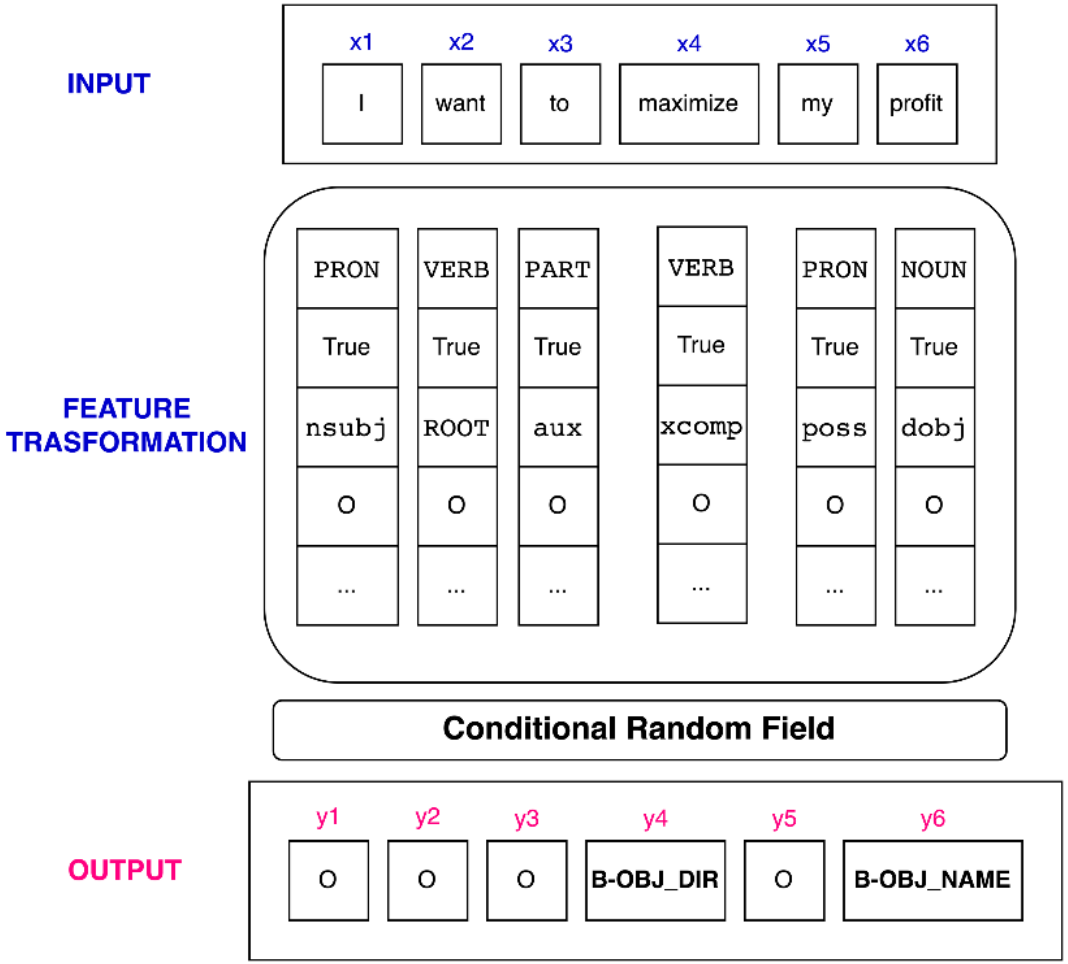
$$w^* = \arg\min_{w}\ L(w, D) + C\,\frac{1}{2}||w||^2 \tag{5}$$

Here, w is the weight vector and C is the regularization parameter.

Lafferty, J.D et. al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ICML 2001

Classical NLP: CRF applied to Ner4Opt

Input → Tokens → Feature Extraction → CRF → OBIE Tags



- In NLP, feature extraction function explores linguistic properties of a token or a group of tokens
- **Grammatical features**: part-of-speech (pos) tagging, dependency parsing, etc.
- **Morphological features**: prefix, suffix and word shape, capitalized, numeric, etc.

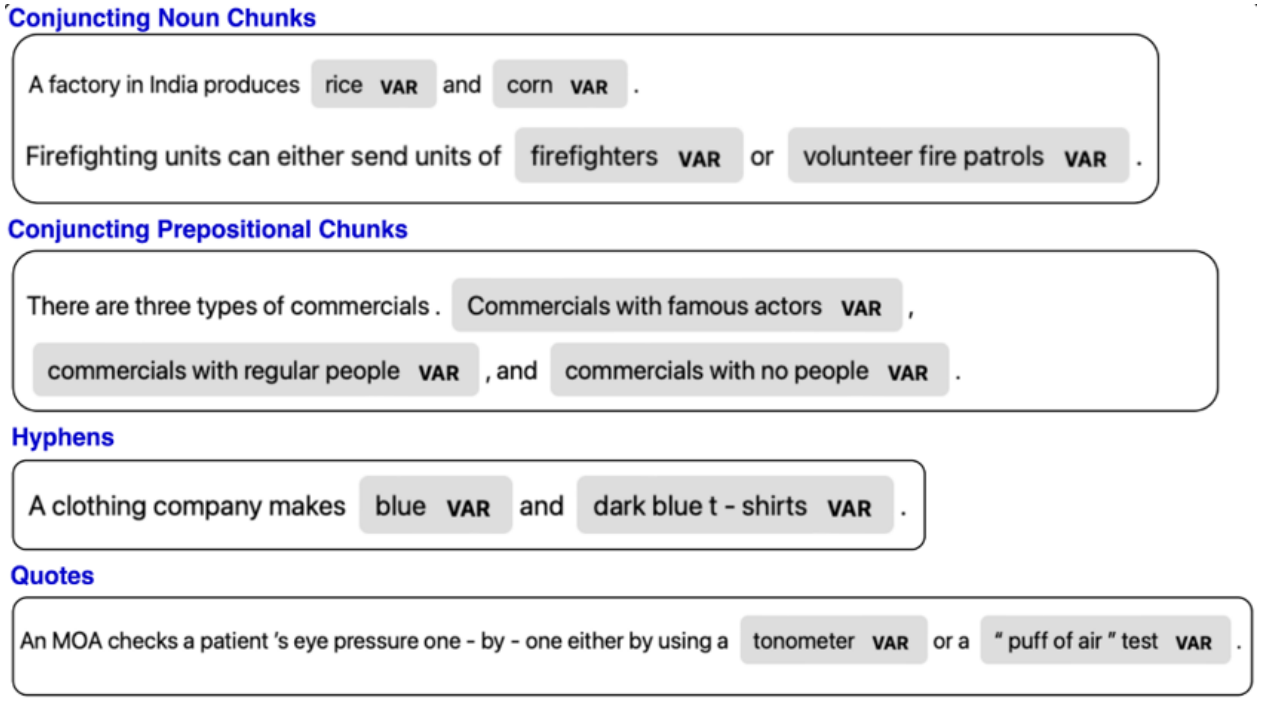
Ratinov, L., Roth, D.: Design challenges and misconceptions in NER, CoNLL, 2009

Feature Engineering for Optimization

Gazetteer and Syntactic features

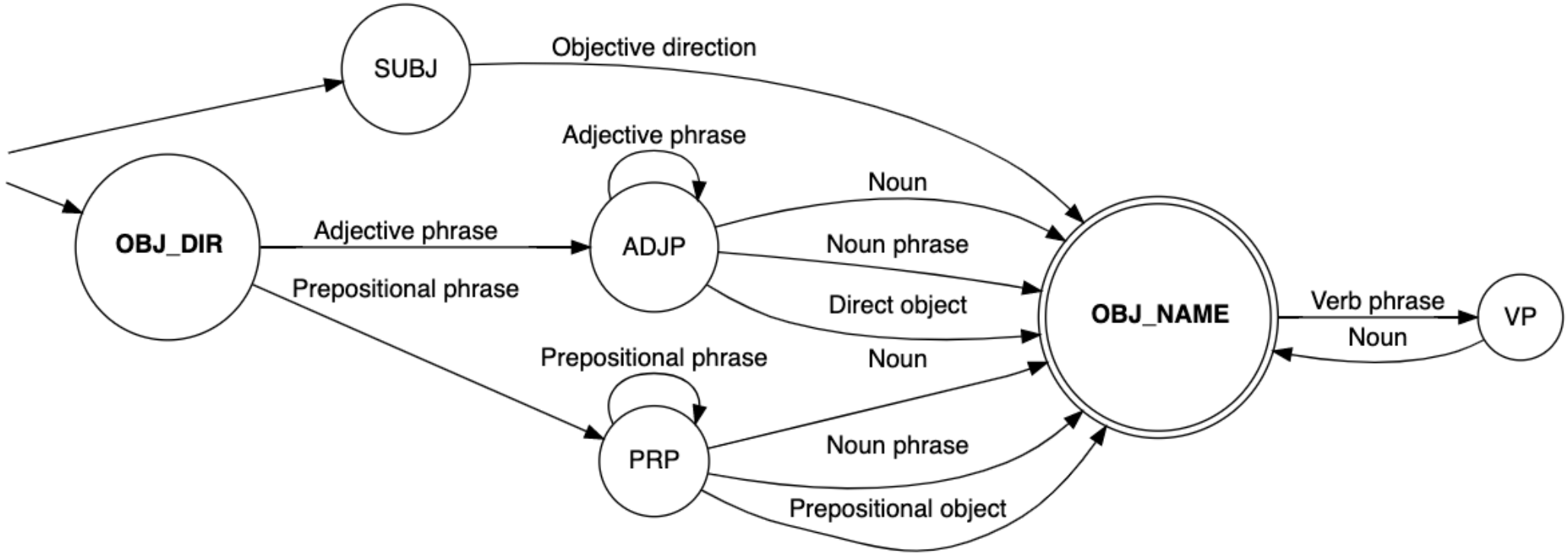
- **Vocabulary features:** gazetteer features serve as lookup tables. Especially useful when the entity class has frequent keywords. **maximize** and **minimize** OBJ_DIR, **at least** and **at most** CONST_DIR

- **Syntactic features:** In linguistics, a **conjunct** is a group of tokens joined together by conjunction or punctuation. VAR and OBJ_NAME entities are associated with unique syntactical properties in the form of conjuncts, noun phrases and propositional phrases, etc.



Regular Automaton for Name Extraction

Extracting the Objective Name



profit SUBJ to be maximized OBJ_DIR

maximize OBJ_DIR the total monthly ADJP profit NOUN

- **Contextual features:** extract left and right context of window size w
- **Constituent parsing**, word-frequency etc.

Modern NLP

Feature Engineering to Feature Learning

- So far, only considered classical methods based on feature extraction and manual feature engineering. This helps us establish a **baseline performance**.
- The challenger to this baseline is motivated by the **recent advances in NLP**, offering advantages over traditional techniques.
- Specifically, **deep neural networks** alleviate the need for manual feature extraction.
- Not only saves a significant amount of but offers more **robust behavior.**
- Moreover, the **nonlinearity in the activation functions** enables learning complex features and dependencies from the labeled training data.

Modern NLP

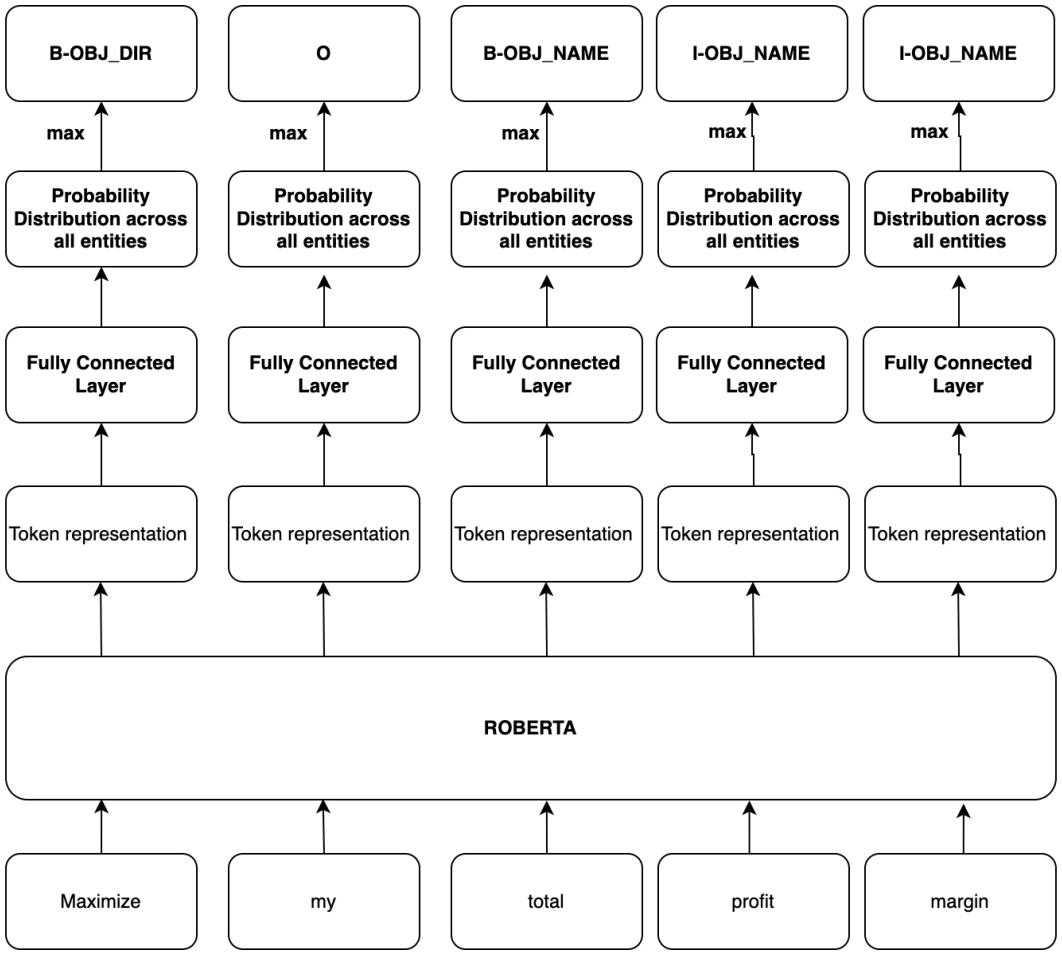
Feature Engineering to Feature Learning

- In practice, Ner4Opt problems require modeling **long-range text dependencies**.
- When operating on the long-range, **recurrent architectures** are known to struggle with vanishing and exploding gradients.
- As a remedy, most recent works rely on the **Transformers architecture** that solve the long-range problem by replacing the recurrent component with the attention mechanism.
- There are many variants of this architecture, and here, we consider distinct flavors based on **RoBERTa** to generate the feature embeddings.

Vaswani et. al.: Attention is all you need, NeurIPS 2017
Liu et. al.: Roberta: A robustly optimized bert pretraining approach, 2019

Formulate Ner4Opt as Token Classification

Use BERT-style models as encoders



Maximize my total profit margin

- ❑ **Token classification** problem with encoders
- ❑ Roberta embeddings with **1024** dimensions
- ❑ A fully-connected layer of size 1024 learns to map token level embeddings into named-entity-labels
- ❑ Followed by **softmax activation function** to output dimension of 1 x 13
- ❑ Minimize training loss with **cross-entropy loss**

Fine-Tuning with Optimization Corpora

Improving LLMs for domain-specific Ner4Opt

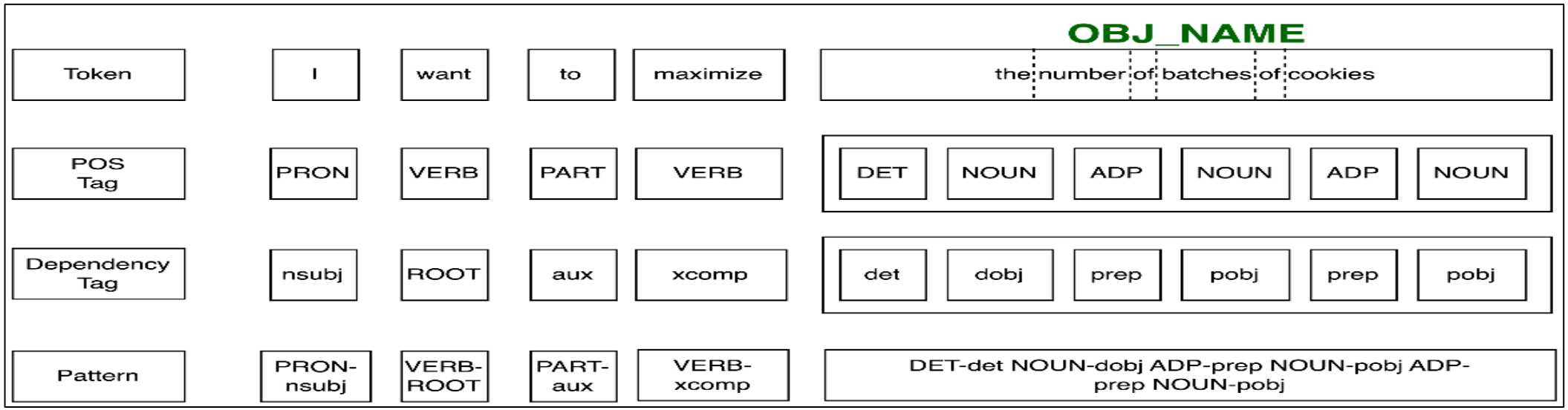
- LLMs, such as BERT, RoBERTa, GPT, are pretrained on **non-domain specific text** for good downstream performance on language-oriented tasks
- For domain specific tasks, performance can be improved using **domain specific corpora** to fine-tune pre-trained models
- Convex optimization, linear programming, game theory books, course notes on optimization from Open Optimization Platform
- Our work is the first approach to fine-tune with optimization corpora using **Masked Language Modelling** with 15% words are random, replace 80% with MAST token, 10% with random, and the remaining 10% with the original word

Howard J., Ruder, S.: Universal language model fine-tuning for text classification, 2018

Data Augmentation

Up-Sampling Infrequent Patterns

- Distribution of classes is balanced. However, **lexical features** exhibit popular traits with infrequent features
- **Example:** objective is maximize/minimize but sometimes as adjective, cost to be minimal
- Challenge is to **find infrequent feature** without manual inspection: Combine POS+DEP Tags



Dealing with Disambiguation

Is it a variable or objective variable?

A doctor can prescribe two types of medication for high glucose levels , a diabetic pill VAR and a diabetic shot VAR . Per dose , diabetic pill VAR delivers 1 PARAM unit of glucose reducing medicine and 2 PARAM units of blood pressure reducing medicine OBJ_NAME . Per dose , a diabetic shot VAR delivers 2 PARAM units of glucose reducing medicine and 3 PARAM units of blood pressure reducing medicine OBJ_NAME . In addition , diabetic pills VAR provide 0.4 PARAM units of stress and the diabetic shot VAR provides 0.9 PARAM units of stress . At most CONST_DIR 20 LIMIT units of stress can be applied over a week and the doctor must deliver at least CONST_DIR 30 LIMIT units of glucose reducing medicine . How many doses of each should be delivered to maximize OBJ_DIR the amount of blood pressure reducing medicine OBJ_NAME delivaered to the patient ?

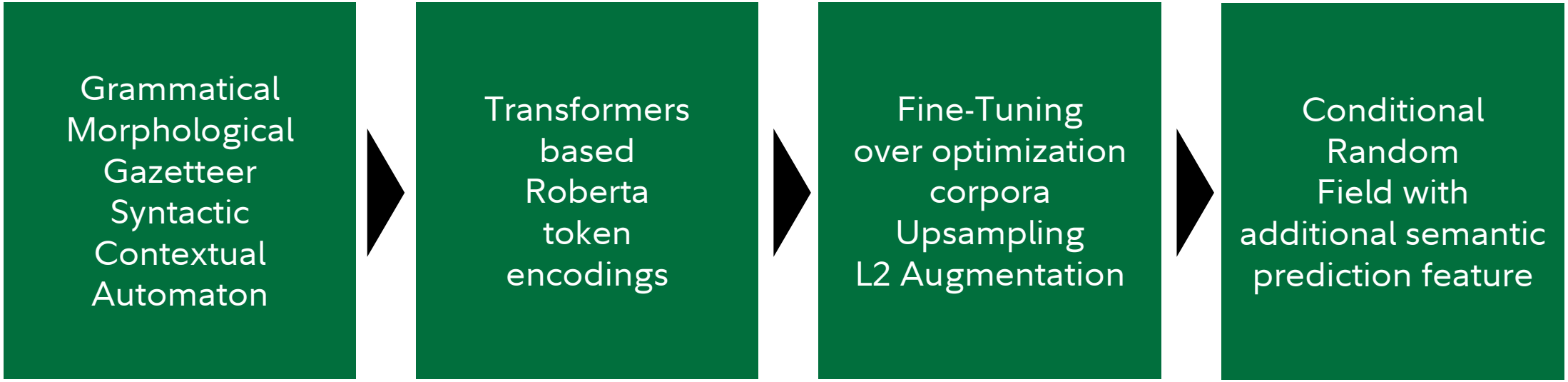
Apply L2 Augmentation

Hybrid Modeling

Feature Engineering + Feature Learning

Feature engineering might be brittle but helps build apriori information

Feature learning brings semantic representations but struggles with long-range dependency



Numerical Results

Effectiveness of the Ner4Opt Solution

Post-mortem and ChatGPT

Experiments

Research Questions

1. What is the baseline classical performance and does feature engineering help?
2. How do modern NLP perform, do we improve over the state-of-the-art?
3. Does the hybrid model perform better than its counterparts in isolation?
4. Where does Ner4Opt fail and how about ChatGPT?

Experiments

Data & Experimental Setup

Statistic	Value
Dataset size	1101
Train set size	713
Dev set size	99
Test set size (not available)	289
Number of entity types	6
Number of VAR entities	5299
Number of PARAM entities	4113
Number of LIMIT entities	2064
Number of CONST_DIR entities	1877
Number of OBJ_DIR entities	813
Number of OBJ_NAME entities	2391

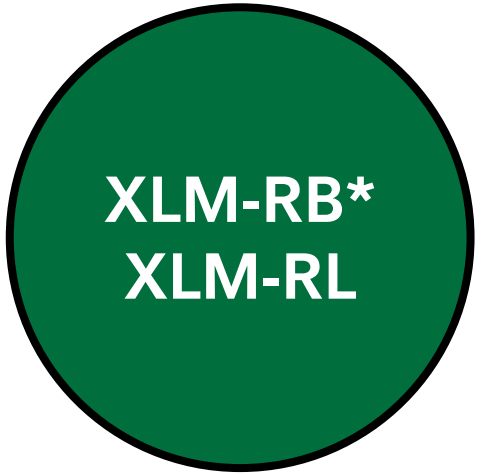
- **Optimization word problems** released as part of NeurIPS'22 NL4Opt Workshop. 1101 optimization instances with annotated entities. 15 annotators
- **Source Domain:** advertising, investment, sales
- **Target Domain:** production, science, transportation
  - Training dataset only comes from Source domain
  - Test and Dev set comes from Source and Target
- **Libraries**: HuggingFace transformers, Simple transformers, SpaCy, sklearn-crf
- Limited hyperparameter tuning to avoid over-fitting

Ramamonjison et. al., NL4Opt Competition: Formulating Optimization Problems Based on Natural Language Descriptions

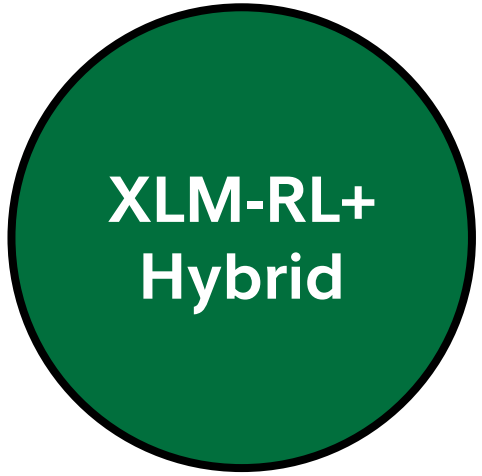
Experiments

Comparisons

Classical based on grammatical and morphological features, plus with hand-crafted gazetteer, syntactic, and contextual features.



The state-of-the-art method* based on XML-Roberta Base and its Large variant



Our optimization fined tuned XML-RL+ and
Hybrid method with feature engineering and learning

* Ramamonjison et. al. Augmenting operations research with auto-formulation of optimization models from problem descriptions, EMNLP, 2022

Experiments

Q1: What is baseline classical performance and does feature engineering help?

Method	CONST_DIR		LIMIT		OBJ_DIR		OBJ_NAME		PARAM		VAR		Average
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	Micro F1
Classical	0.956	0.854	0.904	**0.954**	0.979	0.929	0.649	0.353	0.958	0.916	0.795	0.714	0.816
Classical+	**0.960**	0.858	0.931	0.942	**0.990**	0.970	0.726	0.544	0.953	0.935	0.823	0.787	0.853

$$F1 = \frac{2 * P * R}{P + R}$$

- **Classical+ jumps from 0.81 to 0.85** by hand-crafted gazetteer, syntactic and contextual features
- Feature engineering focus on CONST_DIR and OBJ_DIR which improves
- Classical reports 0.90+ P and 0.85+ R except OBJ_NAME and VAR (ambiguity and long range)

Experiments

Q2: What is the performance of Modern NLP?

Method	CONST_DIR		LIMIT		OBJ_DIR		OBJ_NAME		PARAM		VAR		Average
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	Micro F1
Classical	0.956	0.854	0.904	**0.954**	0.979	0.929	0.649	0.353	0.958	0.916	0.795	0.714	0.816
Classical+	**0.960**	0.858	0.931	0.942	**0.990**	0.970	0.726	0.544	0.953	0.935	0.823	0.787	0.853
Xlm-Rb [51]	0.887	0.897	0.965	0.950	0.949	0.999	0.617	0.469	0.960	0.969	0.909	0.932	0.888
Xlm-Rl	0.930	0.897	0.979	0.938	0.979	0.989	0.606	0.512	0.963	0.985	0.899	0.938	0.893

- **Modern NLP improves over the Classical from 0.81 to 0.88**
- Slight gains when switching to larger models
- Multilingual training of XML is not beneficial for Ner4Opt (compared to RoBERTa)

Experiments

Q3: What the impact of optimization fine-tuning?

Method	CONST_DIR		LIMIT		OBJ_DIR		OBJ_NAME		PARAM		VAR		Average
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	Micro F1
Classical	0.956	0.854	0.904	**0.954**	0.979	0.929	0.649	0.353	0.958	0.916	0.795	0.714	0.816
Classical+	**0.960**	0.858	0.931	0.942	**0.990**	0.970	0.726	0.544	0.953	0.935	0.823	0.787	0.853
Xlm-Rb [51]	0.887	0.897	0.965	0.950	0.949	0.999	0.617	0.469	0.960	0.969	0.909	0.932	0.888
Xlm-Rl	0.930	0.897	0.979	0.938	0.979	0.989	0.606	0.512	0.963	0.985	0.899	0.938	0.893
Xlm-Rl+	0.901	0.897	**0.987**	0.953	0.989	0.999	0.665	0.583	**0.971**	**0.989**	0.918	0.946	0.907

- **Our XLM-RL+ improves with optimization fine-tuning**
- Encouraging result with only a few textbooks over large training corpora
- While higher average score, modern NLP does not improve P/R in every class

Experiments

Q3: What is the performance of Hybrid solutions?

Method	CONST_DIR		LIMIT		OBJ_DIR		OBJ_NAME		PARAM		VAR		Average Micro F1
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	
Classical	0.956	0.854	0.904	**0.954**	0.979	0.929	0.649	0.353	0.958	0.916	0.795	0.714	0.816
Classical+	**0.960**	0.858	0.931	0.942	**0.990**	0.970	0.726	0.544	0.953	0.935	0.823	0.787	0.853
Xlm-Rb [51]	0.887	0.897	0.965	0.950	0.949	0.999	0.617	0.469	0.960	0.969	0.909	0.932	0.888
Xlm-Rl	0.930	0.897	0.979	0.938	0.979	0.989	0.606	0.512	0.963	0.985	0.899	0.938	0.893
Xlm-Rl+	0.901	0.897	**0.987**	0.953	0.989	0.999	0.665	0.583	**0.971**	**0.989**	0.918	0.946	0.907
Hybrid	0.946	0.890	0.980	0.942	**0.990**	**1.000**	**0.730**	**0.668**	0.957	0.983	**0.935**	**0.953**	**0.919**

- **Our Hybrid achieves the best performance 0.919**
- Best performance in most / hardest classes

Post-Mortem

Q4: Where does Ner4Opt solution fails?

```
⟹ How many of each type of donut should be bought in order to
  maximize the total monthly profit OBJ_NAME?
⟹ How many of each type of transportation should the company
  schedule to move their lumber to minimize
  the total cost OBJ_NAME?
⟹ How many of each should the pharmaceutical manufacturing
  plant make to minimize
  the total number of minutes needed OBJ_NAME?
```

- **Conflicting token span** in annotation entities between training and dev sets
- Similar inconsistencies for all classes. Even human annotators cannot agree
- **Aleatoric uncertainty** stemming from data, difficult to address

Post-Mortem

Q4: Why not just use ChatGPT?

KA

Given an optimization problem below, tell me what the

1. Objective is
2. Variables are
3. Constraints are

An consumer electronics business needs to determine the level of production of its two hottest video game consoles , which are the PX7 and Silent X , ahead of the holiday season . Making one PX7 console requires 3 hours of labor and yields a $ 40 profit . On the other hand , one Silent X console can be produced in 7 hours and offers a greater profit of $ 75 . Given the demand forecast , the business decides to produce at least twice as many PX7 consoles as Silent X ones . In addition , it can spend up to 48 hours a week to manufacture these consoles . Can you help the business determine the number of each console to produced each week to obtain the maximum profit ?

1. Objective: The objective is to maximize the profit obtained from the production of PX7 and Silent X video game consoles.
2. Variables: Let X be the number of PX7 consoles produced each week, and Y be the number of Silent X consoles produced each week.
3. Constraints:

- Each PX7 console requires 3 hours of labor, and each Silent X console requires 7 hours of labor. The total labor hours available per week is 48 hours, so the labor constraint is: 3X + 7Y <= 48.
- The business decides to produce at least twice as many PX7 consoles as Silent X ones. Mathematically, this can be written as X >= 2Y.
- The production quantities X and Y cannot be negative: X >= 0 and Y >= 0.

KA

I have the following problem for you.

Given a linear programming optimization problem description, extract named entities.

Example:

Cautious Asset Investment has a total of $ 150,000 to manage and decides to invest it in money market fund , which yields a 2 % return as well as in foreign bonds , which gives and average rate of return of 10.2 % . Internal policies require PAI to diversify the asset allocation so that the minimum investment in money market fund is 40 % of the total investment . Due to the risk of default of foreign countries , no more than 40 % of the total investment should be allocated to foreign bonds . How much should the Cautious Asset Investment allocate in each asset so as to maximize its average return ?

Named Entities:

1. OBJ_DIR (objective direction): maximize
2. CONST_DIR (constraint direction): [total, no more than, minimum]
3. OBJ_NAME (objective name): average return
4. VAR (variables): [money market fund, foreign bonds]
5. PARAM (parameters): [2 %, 10.2 %]
6. LIMIT (limit) : [$ 150,000; 40 %]

Please extract named entities from the text below

An consumer electronics business needs to determine the level of production of its two hottest video game consoles , which are the PX7 and Silent X , ahead of the holiday season . Making one PX7 console requires 3 hours of labor and yields a $ 40 profit . On the other hand , one Silent X console can be produced in 7 hours and offers a greater profit of $ 75 . Given the demand forecast , the business decides to produce at least twice as many PX7 consoles as Silent X ones . In addition , it can spend up to 48 hours a week to manufacture these consoles . Can you help the business determine the number of each console to produced each week to obtain the maximum profit ?

Named Entities:

1. OBJ_DIR (objective direction): maximum profit
2. CONST_DIR (constraint direction): [at least, up to]
3. OBJ_NAME (objective name): profit
4. VAR (variables): [PX7, Silent X]
5. PARAM (parameters): [3 hours, 7 hours, $40 profit, $75 profit]
6. LIMIT (limit): [48 hours a week, at least twice as many PX7 consoles as Silent X ones]

Concluding Remarks & Future Directions

NLP + Optimization: Toward Automated Modelling Assistants

- ❑ Rich literature for integrating ML + Opt but limited work in **NLP + Opt**
- ❑ **Ner4Opt** is immediately relevant for Opt but also challenging task for NLP (counter-intuitive)
- ❑ Improve over the **best-known solution** and show that **generalization** to new domains is possible
- ❑ Common concern: Expertise required as an entry to barrier (learning models, model-seeker, visualizations, explanations, user hints). Our **differentiator is natural text**
- ❑ **HCI questions** when non-technical users are empowered with Opt over text or audio
- ❑ **Call-to-Action:** Let's break the low annotated data regime to realize LLM-style success

AI Center of Excellence @ Fidelity

Open-Source Software

❑ [NeurIPS'22, CPAIOR'23] NER for Optimization	Ner4Opt	https://github.com/skadio/ner4opt
❑ [IJAIT'21] Recommender Systems	Mab2Rec	https://github.com/fidelity/mab2rec
❑ [AAAI'21] NLP/Text Featurization	TextWiser	https://github.com/fidelity/textwiser
❑ [ICTAI'20] Multi-Armed Bandits	MABWiser	https://github.com/fidelity/mabwiser
❑ [AAAI'22, AI Magazine'23] Sequential Mining	Seq2Pat	https://github.com/fidelity/seq2pat
❑ [CPAIOR'22] Feature Selection	Selective	https://github.com/fidelity/selective
❑ [ICMLA'21] Fairness & Bias Mitigation	Jurity	https://github.com/fidelity/jurity

pip install ner4opt

skadio.github.io