A Hybrid Model for Named Entity Recognition in Optimization Problems

Parag Pravin Dakle, Serdar Kadioglu, Karthik Uppuluri, Regina Politi,
Preethi Raghavan, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy

https://github.com/fidelity

Problem Description

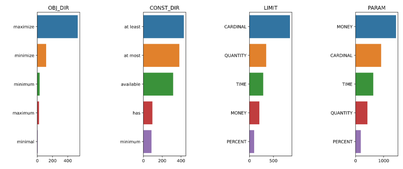
Given an expert formulated optimization problem in natural language, extract six named entities: **CONST_DIR** (constraint direction), **LIMIT** (limit), **OBJ_DIR** (objective direction), **OBJ_NAME** (objective name), **PARAM** (parameter), **VAR** (variable). See example below:

The Notorious Desk company wants to promote a new brand of wine and wants to market it using a total market budget CONST_DIR of $ 87,000 LIMIT . To do so , the company needs to decide how much to allocate on each of its two advertising channels : (1) morning TV show VAR and (2) social media VAR . Each day , it costs the company $ 1,000 PARAM and $ 2000 PARAM to run advertisement spots on morning TV VAR show and social media VAR respectively . The expected daily reach OBJ_NAME , based on past ratings , is 15,000 PARAM viewers for each morning show VAR spot and 30,000 PARAM internet users for a social media spot VAR . The chief marketer knows from her experience that both channels are key to the success of the product launch . She wants to plan at least CONST_DIR 4 LIMIT but no more than CONST_DIR 7 LIMIT morning show VAR spots . In addition , the social media spots VAR needs to be at least CONST_DIR 30 LIMIT due to pricing tier policy . How many times should each of the media channels be used to maximize OBJ_DIR the reach OBJ_NAME of the campaign ?

Data Characteristics

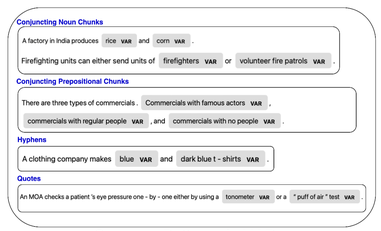
Number of Samples: Training - 713, Dev - 99

Are there any frequently occurring key-phrases or themes in these entities?

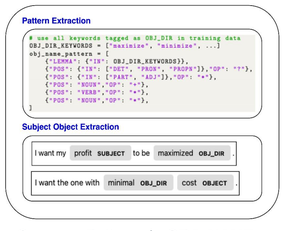


Common key-phrases in **CONST_DIR** and **OBJ_DIR** and Common Themes in **LIMIT** and **PARAM**

Are there any patterns in Structure and Verbiage in these entities?



Common Patterns in **VAR**



Common Patterns in **OBJ_NAME**

Are there any Grammatical/Morphological cues that can be exploited?

- ❑ **OBJ_DIR** is often a verb, with last three characters ending in “ize”
- ❑ **LIMIT** and **PARAM** are some form of numerical entities
- ❑ Both **OBJ_NAME** and **VAR** could be noun chunks / prepositional chunks
- ❑ **OBJ_NAME** sometimes is a direct object / subject in a sentence

The need for Semantic and Contextual Learning (BERT/RoBERTa)

- ❑ All entities except **VAR** have a semantic theme
- ❑ **LIMIT** and **PARAM** are numerical entities which can only be disambiguated by looking at the context in which they appear
- ❑ **OBJ_DIR** and **CONST_DIR** have a directional theme. For instance, phrases like “minimum profit” without any additional context can belong to both classes
- ❑ **OBJ_NAME** often has a concept representing a numeral associated with it (e.g., maximize my profit)

Experimental Protocol

Feature Engineering

- ❑ CRF model exploring basic grammatical and morphological features
- ❑ CRF model exploring grammatical, morphological and engineered features inspired from the Data Characteristics

Feature Learning

- ❑ Token-classification model using RoBERTa large
- ❑ Ensemble of two separate token-classification models one for just OBJ_NAME and VAR and the other for the rest
- ❑ Token-classification model with a modified cost function to optimize for mistakes in OBJ_NAME and VAR
- ❑ Token-classification model using XLM-RoBERTa and curriculum learning
- ❑ Token-classification model using XLM-RoBERTa fine-tuned on Optimization Corpora

Hybrid

- ❑ CRF Model combining best performing Feature Engineering and Feature Learning techniques

Augmentation Strategies

Up sampling via Duplication of in-frequent patterns

- ❑ **OBJ_DIR** is generally a verb (e.g., maximize, minimize) but there are a few examples, where **OBJ_DIR** is also an adjective (e.g., I want the cost to be **minimal**)
- ❑ **VAR** is mostly a Conjuncting noun chunk. Conjuncting prepositional phrases are an infrequent pattern (e.g., He does commercials with famous actors and commercials with regular actors)
- ❑ **OBJ_NAME** is **OBJ_DIR** followed by a noun phrase / prepositional phrase. **OBJ_DIR** followed by multiple prepositional phrases is a rare pattern (e.g., maximize the number of action figures; minimize the number of batches of cookies)

Augmenting Last Two Sentences: In most cases, for **OBJ_NAME** tokens to be tagged correctly it is imperative that the objective is known first**. For** example:

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 ?

Pseudo Label Data generation: Use paraphrase corpora like WordNet and PPDB to generate pseudo label data

Hybrid Model

Feature Engineering

- ❑ Grammatical Features
- ❑ Morphological Features
- ❑ Gazetteer Features
- ❑ Features exploiting syntax and verbiage
- ❑ Features are extracted at each word position and a window around it

Feature Learning

- ❑ Label predictions from a trained RoBERTa transformers model
- ❑ Large variant and the base variants is used for comparison

Conditional Random Field

Selected Results

Model Name	CONST_DIR		LIMIT		OBJ_DIR		OBJ_NAME		PARAM		VAR		Average Micro F1 (Dev)	Average Micro F1 (Test)
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*		
Grammatical and Morphological Features + CRF	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	-
Grammatical, Morphological, Gazetteer, Structural Features + CRF	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	-
RoBERTa Large	0.895	0.902	0.984	0.950	0.990	1.000	0.668	0.597	0.965	0.983	0.916	0.940	0.904	-
[Infrequent Pattern Upsampling] + RoBERTa Large	0.947	0.909	0.984	0.950	0.990	0.990	0.628	0.615	0.961	0.979	0.906	0.947	0.903	-
Pre-trained XLM-RoBERTa Large on Textbooks	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	
[Last Two Sentence Augmentation] + Grammatical, Morphological, Gazetteer, Structural Features + RoBERTa Model Predictions + CRF + RandomSearchCV	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**	**0.920**

Discussions & Observations

Scope for Aleatoric Uncertainty - Similar sequences annotated differently in Train and Dev

How should the bakery operate to maximize OBJ_DIR total profit OBJ_NAME ?

How many of each type of transportation should the company schedule to move their lumber to minimize OBJ_DIR the total cost OBJ_NAME ?

How many of each type of donut should be bought in order to maximize OBJ_DIR the total monthly profit OBJ_NAME ?

If the chemical company needs to make at least CONST_DIR 900 LIMIT au of the acidic liquid and 1200 LIMIT au of the basic liquid per minute OBJ_NAME , how many minutes OBJ_NAME should each reaction be run for to minimize OBJ_DIR the total time OBJ_NAME needed ?

How many of each should the pharmaceutical manufacturing plant make to minimize OBJ_DIR the total number of minutes needed OBJ_NAME ?

Cautious Asset Investment has a total CONST_DIR of $ 150,000 LIMIT to manage and decides to invest it in money market fund VAR , which yields a 2 % PARAM return OBJ_NAME as well as in foreign bonds VAR , which gives and average rate of return OBJ_NAME of 10.2 PARAM % .

To do so , the company needs to decide how much to allocate on each of its two advertising channels : (1) morning TV show VAR and (2) social media VAR . Each day , it costs the company $ 1,000 PARAM and $ 2000 PARAM to run advertisement spots on morning TV VAR show and social media VAR respectively .

Excerpts from Train (green) and Dev (yellow) highlighting annotation inconsistency for similar sequences