Dominating word sets

Distant-CTO

- Matching “Intervention” terms to free-text in clinicaltrials.org
- Direct match and fuzzy span matching using bigram matches lots of items.
- How could dominating word sets (order-free) could help?
    - Identify examples where order-free match could help

Where it might work?

1. Home screening - “home screening” could match “Home-based screening” in in brief summary
2. Topiramate coated tablet - “Topiramate coated tablet” could match “topiramate 100 mg, coated tablet” in detailed description
3. Upper limb rehabilitation - “upper limb rehabilitation” could match “Upper Limb (UL) rehabilitation” in brief summary
4. Cardio pulmonary exercise testing - “cardio-pulmonary exercise testing (CPEX)” could match “cardio-pulmonary exercise (CPEX) testing” in detailed description

Order-free source matches over targets

Where it might work?

1. YESplus workshop - “YESplus workshop” could match “(YESplus) workshop” in brief summary
2. Mindfulness intervention - “Mindfulness intervention” could match “mindfulness meditation intervention” in intervention/treatment table
3. Donor leukocytes - “Donor leukocytes” could match “leukocytes from a donor” in brief summary
4. Ozone injection - “ozone injection” could match “ozone, prolotherapy injection” in the brief summary

Order-free source matches over targets

Where it won’t work?

- Rehabilitation Strength training - The distant supervision source “Rehabilitation strength training” occurs multiple times in the documents (here; sentences) but the word rehabilitation is replaced with either physiotherapy or exercise or exercise therapy.
- Ozone injection - Ozone injection is mentioned as “ozone therapy” in the brief summary section.

Notes on application to Distant-CTO

- Sentence should be considered as a document here (not paragraph)
- Lemmatization should help as well
- Number of false positives might depend on the value of sliding window parameter.
- The work has potential to follow order-free matches.
- But can an order-free match also be dominant word set is a matter of trial?
    - Reason: such order-free distant supervision matches might occur just once or twice in a CTO study where it is mentioned.

Conclusion?

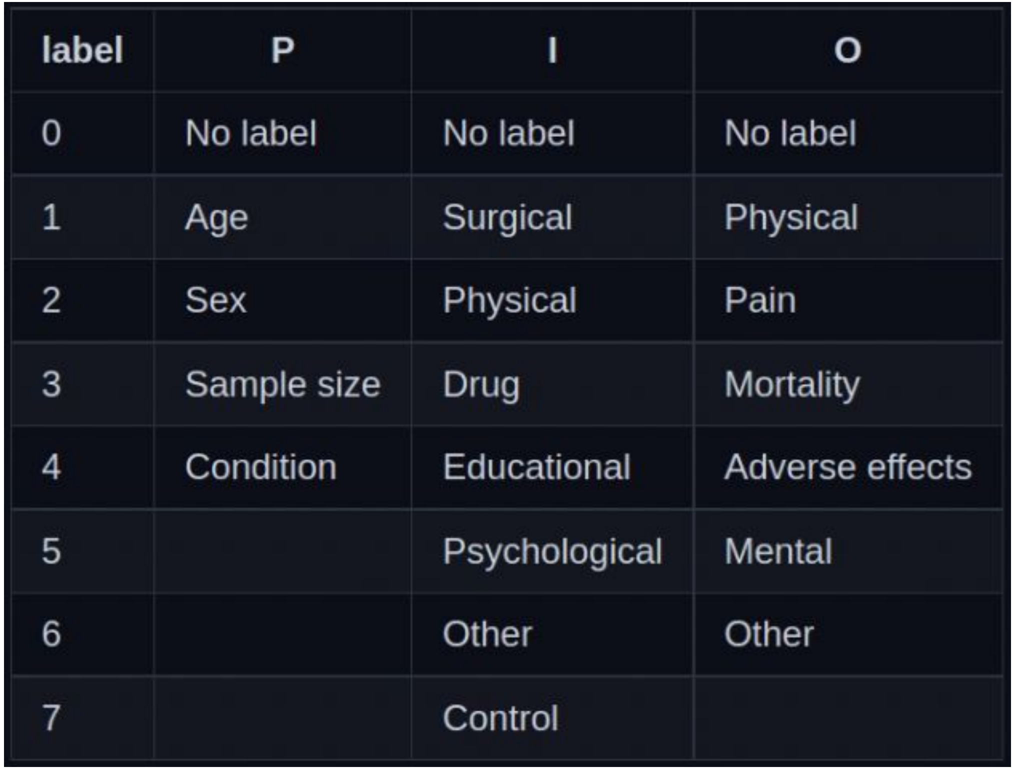
- Might help reduce false negatives during candidate generation, but might also increase false positives.
- How much benefit could it bring? > Quantification of results requires trial and error
- Quantification of results require an actual validation set (manually annotated) to test on.
- The solution might be relevant only for *clinicaltrials.org* (but I might be wrong!)

Further investigation

- **Can order-free matches improve distant/weak labeling of “Intervention” mentions in the text?**
- EBM-PICO training set
- EBM-PICO validation set (to improve the heuristics and parameter optimization for order-free matching)
- What metrics could be measured on the validation set?
    - Coverage, F1-score, TPR, TNR, Optimize on recall
- Train DL/ML models on EBM-PICO training+validation sets
- EBM-PICO gold test set (to evaluate the DL/ML models)

Datasets available

1. EBM-PICO training set
2. EBM-PICO gold test set
3. Physio test set



label	P	I	O
0	No label	No label	No label
1	Age	Surgical	Physical
2	Sex	Physical	Pain
3	Sample size	Drug	Mortality
4	Condition	Educational	Adverse effects
5		Psychological	Mental
6		Other	Other
7		Control	

Experimental design: Candidate Generation

Distant Supervision (Intervention)	**Matching**	**Coverage**	**F1 score**	**TPR**	**TNR**
Distant supervision from CTO	Order-preserving matching (OM)				
	Dynamic programming based loose matching (DP)				
	Relevant bigram matching (RB)				
	Order-free matching (OF)				

Experimental design: Candidate Generation

Distant Supervision (Intervention)	**Matching**	**Coverage**	**F1 score**	**TPR**	**TNR**
Distant supervision from CTO	OM + DP				
	OM + DP + RB				
	OM + DP + RB + OF				

Experimental design: Candidate Generation

Distant Supervision (Intervention)	**Matching**	**Coverage**	**F1 score**	**TPR**	**TNR**
Distant supervision from CTO + external ontologies	Direct matching (DM)				
	Dynamic programming based loose matching (DP)				
	Relevant bigram matching (RB)				
	Order-free matching (OF)				

Experimental design: Candidate Generation

Distant Supervision (Intervention)	**Matching**	**Coverage**	**F1 score**	**TPR**	**TNR**
Distant supervision from CTO + external ontologies	DM + DP				
	DM + DP + RB				
	DM + DP + RB + OF				

Experimental design: Model training

- Labeling functions: Participant - disease, Intervention, Outcomes
- Label all the previously-mentioned datasets
- Training discriminative models with the generated candidates using different labeling functions.
- Tracking metrics: Precision, Recall and F1.
- Most important metric is Recall because SRs are recall-oriented task.

Resources

- EBM-PICO training, validation and test sets - available as json
- Distant supervision (CTP) labeling source - available as json
- External ontologies labeling source - available in csv, tsv files
- Implementation of dynamic programming based loose matching - on Github

Input

interventions.txt: line X : “Topiramate”, “coated”, “tablet”

Test_ebm_… : line Y : “The”, “objective”, “is”, “to”, “confirm”, “if”, “two”, “formulations”, “of”, “topiramate”, “100”, “mg,”, “coated”, “tablet,”, “are”, “bioequivalent,”, “after”, “oral,”, “single-dose”, “administration”, “under”, “fasting”, “conditions”.

Output

“The”, “objective”, “is”, “to”, “confirm”, “if”, “two”, “formulations”, “of”, “**topiramate**”, “100”, “mg,”, “**coated**”, “**tablet**,”, “are”, “bioequivalent,”, “after”, “oral,”, “single-dose”, “administration”, “under”, “fasting”, “conditions”.