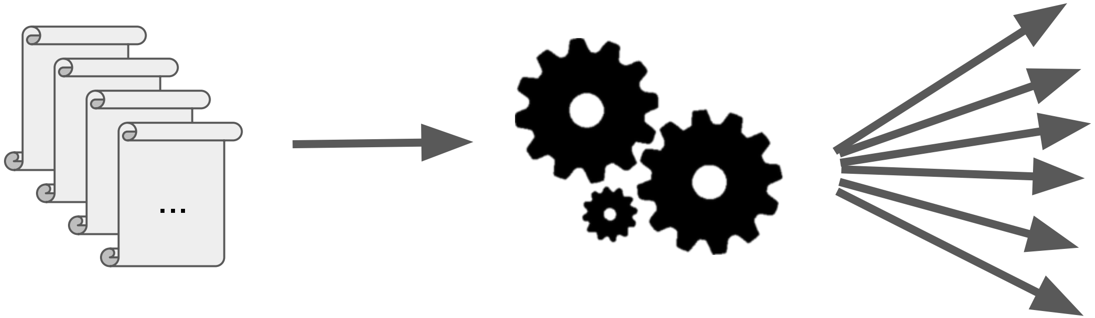
On our submission to the Covid-19 MLIA Eval Task 1

SWLab@uniCA

Sandro Gabriele Tiddia - Diego Marcia - Manuela Sanguinetti - Maurizio Atzori

Task: extracting text spans from an untagged corpus



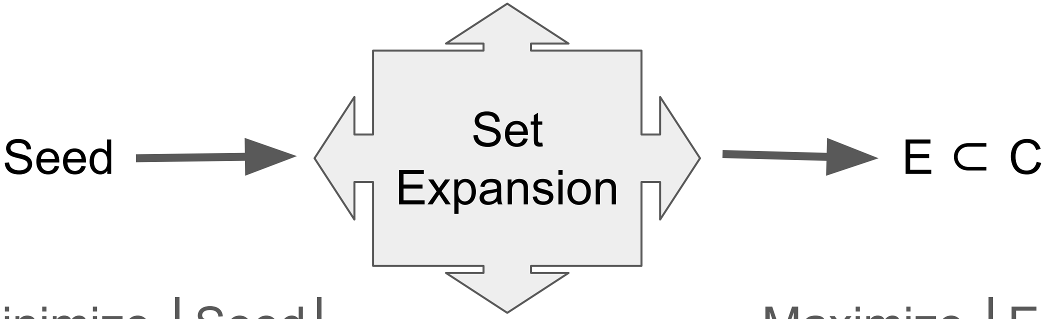
BRAT annotations

- *{drug-trt}*
- *{sosy-dis}*
- *{findings}*
- *{tests}*
- *{behavior}*
- *{legal-reg}*

Set Expansion

$C \in \{ \text{drug-trt} ; \text{sosy-dis} ; \text{findings} ; \text{tests} ; \text{behavior} ; \text{legal-reg} \}$

$\text{Seed} = \{ s_0 , \ldots , s_n \} \subset C$



Seed → Set Expansion → $E \subset C$

Minimize $|\text{Seed}|$

Maximize $|E|$

OKgraph

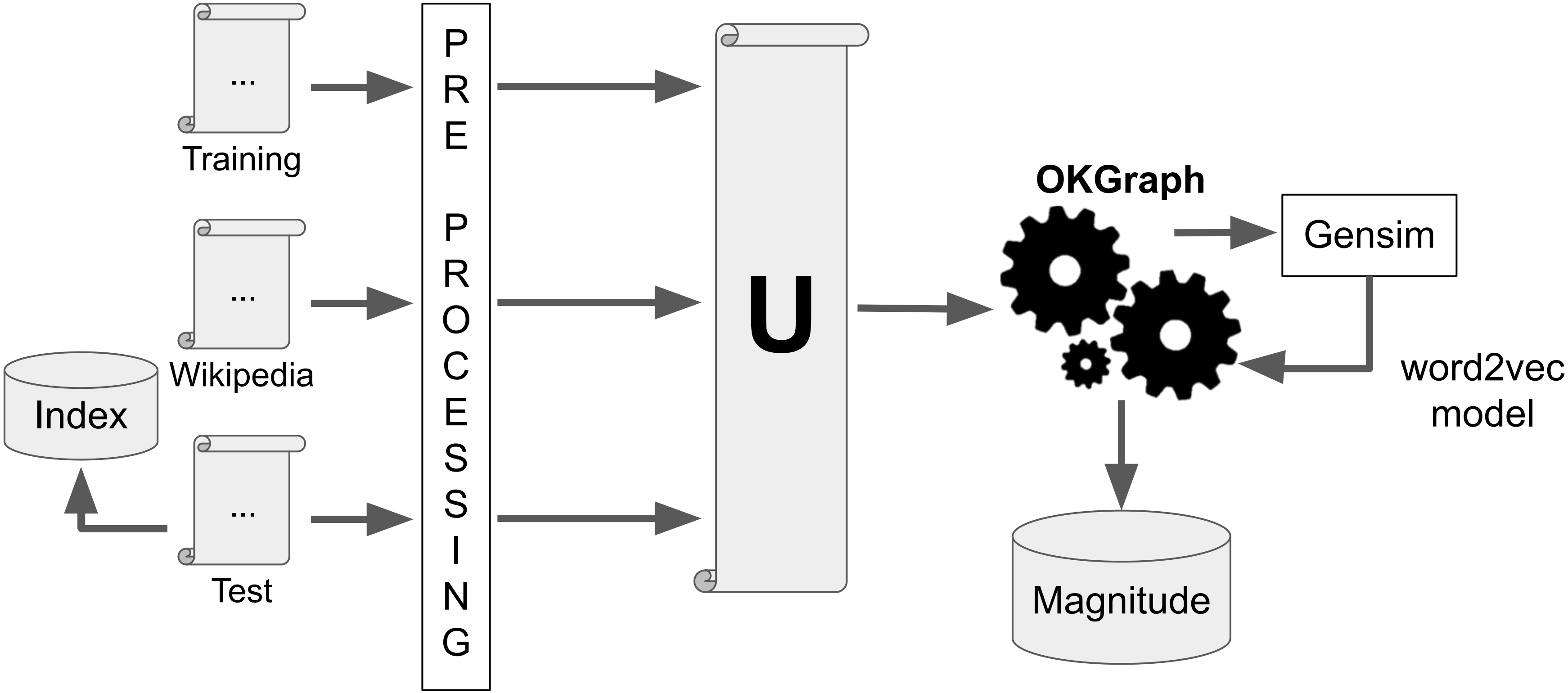
- relation_expansion
- relation_labeling
- set_labeling
- set_expansion
  - Scored tree expansion
  - k-nearest neighbours
  - k-nearest neighbours with centroid boost
    - Vector representation of words

Class decomposition

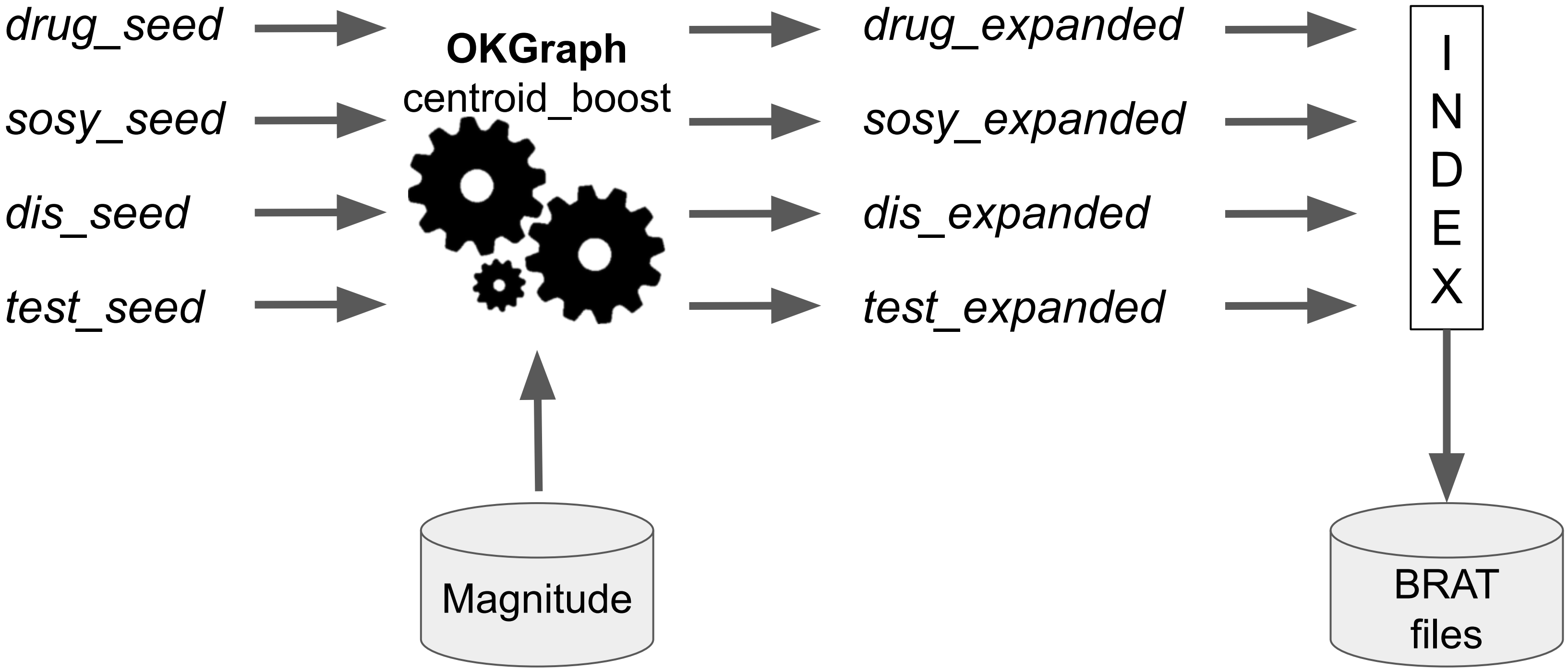
- drug-trt
  - drugs : *{ improvac ; pemetrexed lilly ; protopic }*
  - ~~treatments : { … }~~
- sosy-dis
  - signs or symptoms : *{ breathing difficulty ; disorientation ; blindness }*
  - diseases : *{ tardive dyskinesia ; diabetes mellitus ; cardiomyopathy }*
- tests : *{ screening ; ct scan ; mammography }*
- ~~legal-reg : { … }~~

Findings and *behavior* ignored, *treatments* and *legal-reg* dismissed after first tests

Pipeline: training



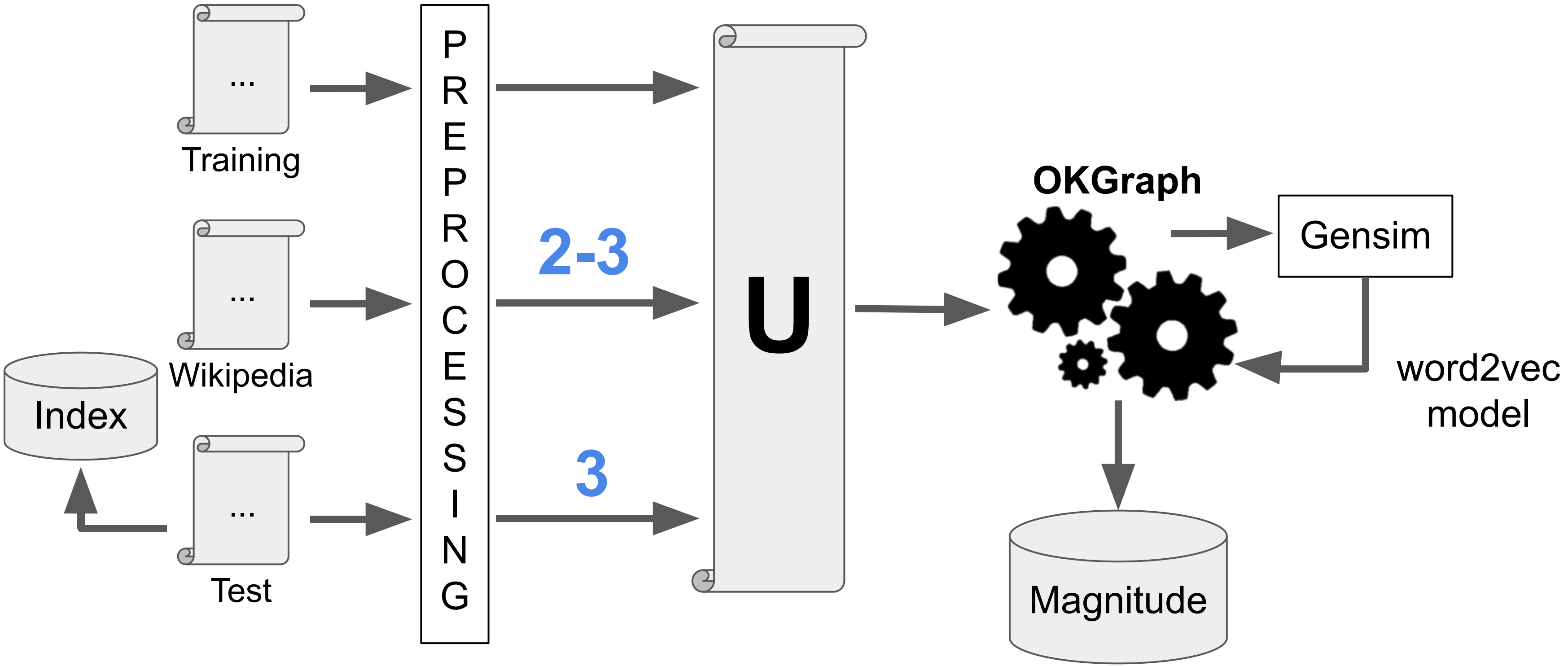
Pipeline: k-nearest neighbours (k=200)



Three runs / Two Languages

English	**Italian**
1. Model trained with the training corpus	
2. Training + 1GB from Wikipedia dump	1. Training + 1GB from Wikipedia dump
3. Training + Wiki + test corpus	2. Training + Wiki + test corpus

Three runs (continued)



Scores

- 170 predictions
- Precision .004 for class *sosy*
- Precision .105 for class *tests*
- Precision .000 everywhere else

Rover			
Run 1	3657-ab.ann	sosy-dis	Mood swings
Run 2	3706-aa.ann *	test	clinical examination
Run 2	3708-aa.ann	sosy-dis	Breathing Difficulties
Run 3	3657-ab.ann	sosy-dis	Mood swings
Run 3	3706-aa.ann *	test	clinical examination

* One occurrence in ROVER, two in our submission

- 8 to 9 total predictions
- Precision .001 for class *sosy* - 2
- Precision .000 everywhere else

Gold Standard			
Run 2	3708-aa.ann	sosy-dis	Breathing Difficulties

Index lookup limit

- Indexing was performed via an external library (Whoosh)
- Default behaviour limits the number of retrieved matches

Run	Predictions	sosy (n=1173)	tests (46)	overall
1	49	.2632	.0333	.1224
2	71	.2400	.0217	.0986
3	63	.2174	.0250	.0952

- Values for precision
- Evaluated on the gold standard

* Code and data in repo are still at “buggy” version

Better seed sets

- The entities in the seed sets are representative of the six categories, but not strictly related to Covid-19
- Post-evaluation experiments with 21 different seeds (run 3, category *sosy-dis*)

- One tailored seed set for diseases: *{ sars ; coronavirus ; pneumonia }*
- 20 random sets from its expansion

- Tailored set achieved the best recall and F-score (+1~2 orders of magnitude!)
- Random sets very similar in terms of F-score (σ^2=0.0004 vs σ^2=0.0089)

Tuning k-nearest neighbour

- For our submission, the *k* parameter has been arbitrarily set to 200
- We performed some post-evaluation experiments, with the same setting from run #3, restricted to class *diseases*, with the tailored seed set

k	Annotations	Precision	Recall	F-score
20	306	**.8105**	.2166	.3418
175	668	.4895	.2769	**.3537**
200	675	.4844	.2769	.3524
500	729	.4540	**.2798**	.3462

Conclusions and Next Steps

- Our submission performed poorly wrt both the rover and the gold standard
- Correction of an ill-valued parameter improved the result…
- … But not for all classes!
- Selecting a tailored seed set is crucial for recall, but we don’t have a method for automating its selection yet
- Tuning of the *k* parameter for nearest neighbour, so to find a balance between precision and recall, should be taken into consideration

Results suggest that Set Expansion, in the context of an Information Extraction task, tends to produce higher-Precision and lower-Recall results.