Weak Supervision for Cross-Sentence Relation Extraction

Felipe Arias[1], Alex Ratner[2], Christopher Ré[2]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL

[2]Department of Computer Science, Stanford University, Stanford, CA

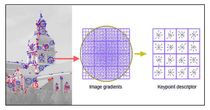
Motivation

The traditional ML pipeline has a new bottleneck.

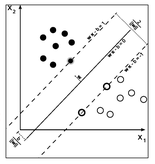
Collect Raw Data → Create a Training Set → ~~Engineer Features~~ → Fit a Model

True

False

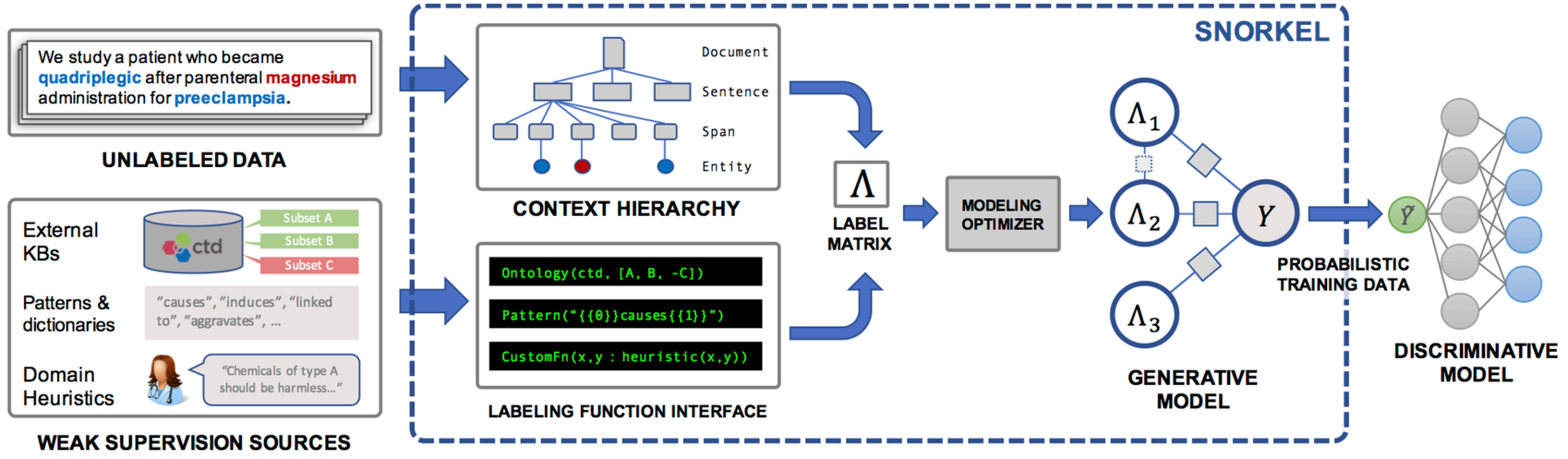


Now we use deep learning, but it's data hungry!



Snorkel, a system for creating training data, has improved on distant supervision/heuristic baselines and performed comparably to hand-labeled datasets. This project aims to extend Snorkel's label generating abilities to include cross-sentence relation extraction.

Snorkel



Candidate Extraction

Ex: Extracting all person-car-color relations across three sentences or less in a document:

	Sentence 1	Sentence 2	Sentence 3
Entity 1	Han	Joe, Ben	Carl
Entity 2	Audi	Ford, Fiat	
Entity 3	Blue	Purple	Gray

In order to find all 3-ary relations that include entity mentions from the first sentence we do the following:

C_1 = (Han) × (Audi, Ford, Fiat) × (Blue, Purple, Gray)
C_2 = (Joe, Ben, Carl) × (Audi) × (Blue, Purple, Gray)
C_3 = (Joe, Ben, Carl) × (Ford, Fiat) × (Blue)

The candidates for the first sentence are then $\bigcup_{i=1}^{3} C_i$.

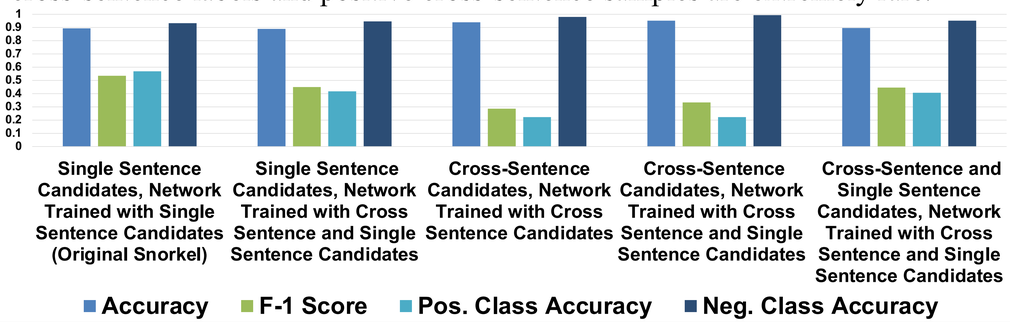
Continue by iterating over the remaining sentences and updating the queue by discarding the first sentence and adding the next.

Queue

Han	Joe, Ben	Carl	
Audi	Ford, Fiat		Jeep
Blue	Purple	Gray	Black

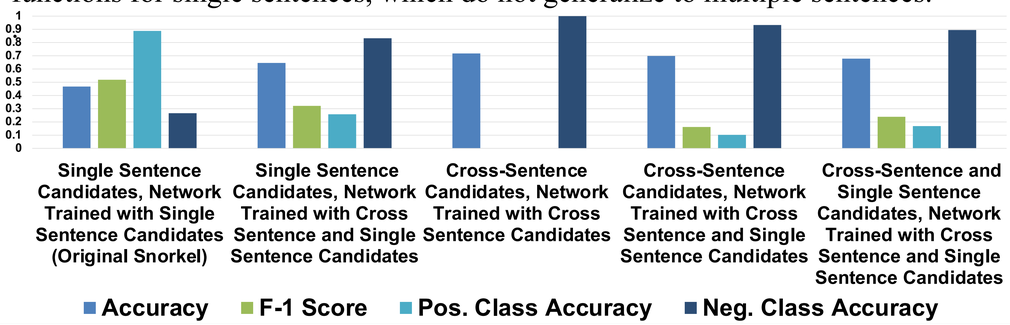
Extracting Spouse Relations

Below are performance metrics for various Snorkel configurations tested on a spouse dataset. We found this dataset to be undesirable as it does not come with cross-sentence labels and positive cross-sentence samples are extremely rare.



Extracting Chemical – Disease Relations

Below are performance metrics for various Snorkel configurations tested on a chemical-disease dataset. As with the other example, all models used labeling functions for single sentences, which do not generalize to multiple sentences.



Conclusion

Although the preliminary models do not always perform at the same level as the single sentence (original) version of Snorkel, the potential for performance improvement and usefulness is evident. Cross-sentence relation extraction enables more complexity when writing labeling functions and deciding on a discriminative model, which could, in turn, permit better performance. In addition, grid search for hyperparameter optimization and an increase in the number of training epochs would benefit the models and result in a superior comparison.

Acknowledgements

- A special thanks to Alex Ratner, Christopher Ré, and everyone in Hazy Research for letting me work with them and supporting me in my research.
- Made possible and funded by SURF.

More Information

- snorkel.stanford.edu (Source code and tutorials)
- A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. VLDB.