A Comparison of Weak Supervision Methods for KBC

A. Soni[1], D. Viswanathan[2], N. Pachaiyappan[2], S. Natarajan[2]

[1]Swarthmore College, [2]Indiana University

Introduction

Creating gold-standard training sets can be prohibilitively expensive for tasks such as KBC. Commonly, researchers use **weak** or **distant supervision** techniques to more cheaply produce large training sets. This work analyzes three such approaches for producing so called silver-standard examples.

Central Question

Which weak supervision techniques provide the best basis for learning accurate models and scale appropriately with the KBP task?

Results and Discussion

Relation	KWS	CDS	EDS
age	500	0	0
parents	413	128	782
spouse	1533	346	403
siblings	773	43	325
foundBy	148	239	2207
countryHQ	21	168	1715

Number of Examples Across each person relation, KWS tends to outperforms despite using only 5% of documents. One relation, *age*, could not be mapped to any known db, yielding 0 DS examples. EDS utilizes more seed entries from db than CDS.

Relation	KWS	CDS	EDS
parents	**0.68**	0.62	0.49
spouse	**0.81**	0.46	0.53
siblings	**0.69**	0.52	0.58
foundBy	0.60	**0.72**	0.63
countryHQ	0.58	0.69	**0.69**

Area Under ROC Curve RDN models trained using each of the three training sets. Results are across 5 runs. KWS outperforms across all person relations, but struggles with organizations where rules are more difficult to encode. Gold examples help KWS equal DS in *countryHQ* relation.

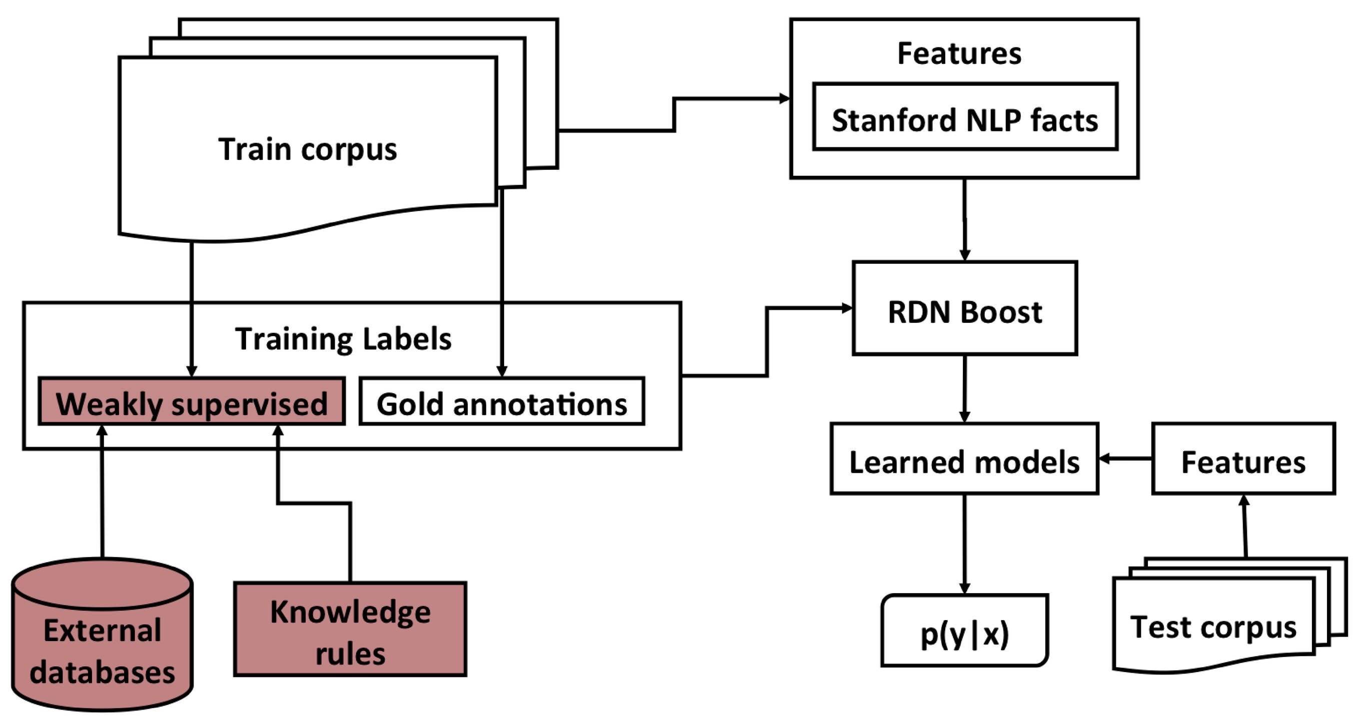
Methods

Weak Supervision Approaches

Distant Supervision refers to an external database as a source of seed examples [1] (NELL, Wikipedia Infoboxes, and Freebase).

- *Corpus Distant Supervision* (**CDS**) - map external db entries to sentences in a corpus native to the test domain (e.g., TAC KBP newswire articles). Emphasizes matching learned models to test space.
- *External-Text Distant Supervision* (**EDS**)- map db entries to sentences in an external text (Wikipedia articles). Emphasizes utilization of seed examples.
- *Knowledge-Based Weak Supervision* (**KWS**) - encode "world knowledge" of domain experts [2] e.g., in FOL rules. Apply rules (via MLNs) to corpus to generate positive training examples.

Pipeline



To evaluate, we trained Relational Dependency Networks (RDN) [4] using training examples from each of CDS, EDS, and KWS (red above). TAC KBP 2014 corpus was used for training while 2015 for testing. We considered 6 TAC KBP relations (see results) representing persons and organizations.

Weight	KWS Rule (MLN Clause)
1.0	*entityType(a, "PER"), entityType(b, "NUM"), nextWord(a, c), word(c, ","), nextWord(c, b) → age(a, b)*
0.8	*entityType(a, "PER"), entityType(b, "PER"), nextLemma(a, "mother") → parents(a, b)*
0.6	*entityType(a, "PER"), entityType(b, "PER"), lemmaBetween(a, b, "husband") → spouse(a, b)*
1.0	*entityType(a, "ORG"), entityType(b, "PER"), prevPrevLemma(b, "found"), prevLemma(b, "by") → foundedBy(a, b)*

Conclusions

Knowledge-based WS is viable alternative/complement to the popular distant supervision approaches

- can produce more examples and better models
- requires good (formalisible) world knowledge

Distant supervision techniques

- scale better with larger corpus
- yield fewer results; require existing db
- can utilize external texts if (native) training data is unavailable

We acknowledge Deep Exploration and Filtering of Text Program under the Air Force Research Laboratory (AFRL) prime contract and FA8750-12-2-0039

References

[1] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data.

[2] S. Natarajan, J. Picado, T. Khot, K. Kersting, C. Re, and J. Shavlik. 2014. Effectively creating weakly labeled training examples via approximate domain knowledge.

[3] F. Niu, C. Re´, A. Doan, and J. W. Shavlik. 2011. Tuffy: Scaling up statistical inference in Markov logic networks using an RDBMS.

[4] S. Natarajan, T. Khot, K. Kersting, B. Gutmann, and J. Shavlik. 2010. Boosting relational dependency networks.