Crowdsourcing Semantic Label Propagation in Relation Classification

ANCA DUMITRACHE, LORA AROYO
Vrije Universiteit Amsterdam
{anca.dmtrch, l.m.aroyo}@gmail.com

CHRIS WELTY
Google Research
cawelty@gmail.com

Distant Supervision (DS) is a method for annotating sentences with relations by aligning a knowledge base with a text corpus. However, the assumption that, given a triple in the knowledge base, every sentence in the corpus that contains the triple will also contain the relation, generates **a lot of noisy data**.

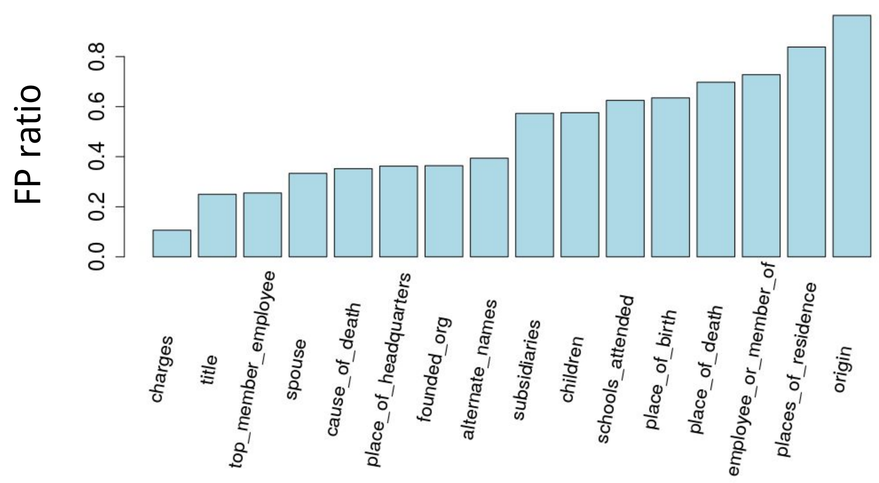
CrowdTruth is a methodology for crowdsourcing ground truth data by harnessing inter-annotator disagreement. Key metric:

- **sentence-relation score (SRS):** ratio of workers that picked the relation in the sentence over all workers, weighted by worker quality

Crowdsourcing for Relation Label Propagation

Crowdsourcing Setup

- 4,100 sentences, annotated with Distant Supervision, split into 2,050 dev & 2,050 test
- 17 relations + no relation
- 15 workers / sentence
- workers are given sentence + term pair, asked to select which relations apply from multiple choice list
- evaluation shows high false positive rate for some relations in Distant Supervision:



Label Propagation (1,2)

- update label of Distant Supervision sentence using the *SRS* of the most similar sentence in the Crowd dev set
- **sentence embedding:** word2vec word embeddings(3), averaged across all words in a sentence(4)
- **similarity function:** cosine similarity
- **label update function:** given sentence *s*, relation *r*, and the original distant supervision label *DS(s, r) ∈ {0, 1}*:

$$DS^*(s,r) = \frac{DS(s,r) + cos\_sim(s,l') \cdot srs(l',r)}{1 + cos\_sim(s,l')}$$

$$l' = argmax_{l \in Crowd\_dev} cos\_sim(l,s)$$

(1) Sterckx et al. “Knowledge base population using semantic label propagation.” Knowledge-Based Systems, 108(C):79–9. 2016.
(2) Xiaojin, Zoubin. “Learning from labeled and unlabeled data with label propagation”. CMU-CALD-02–107, CMU. 2002
(3) Mikolov et al. “Distributed representations of words and phrases and their compositionality”. NIPS. 2013
(4) Sultan et al. “DLS @ CU: Sentence Similarity from Word Alignment and Semantic Vector Composition.” SemEval workshop. 2015.

Enhancing Distant Supervision with CrowdTruth

Relation Extraction Model

- **convolutional neural network**(5)
- **features:**
  - word2vec word embeddings, initialized with pre-trained Google News(6)
  - term pair position embeddings, with random initialization
- **loss function:** sigmoid cross-entropy, computed on continuous labels
- **architecture:**
  - embedding layer, updated during training
  - convolutional layer with sliding window of 2 to 5 words, learning *n-grams*
  - pooling layer, learning *sentence-level features*
  - sigmoid layer for *multi-class multi-label classification*

(5) Nguyen, Thien Huu, and Ralph Grishman. "Relation Extraction: Perspective from Convolutional Neural Networks." VS@ HLT-NAACL. 2015.

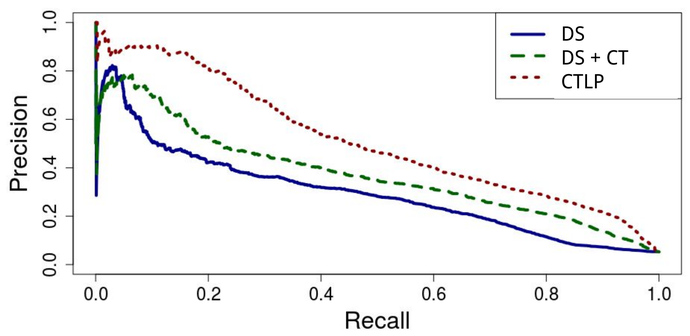
(6) https://code.google.com/archive/p/word2vec/

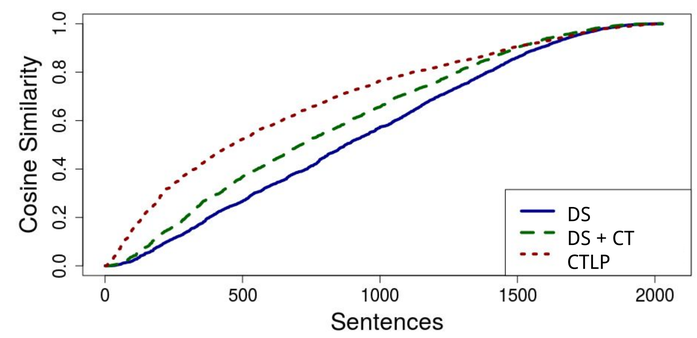
Training Data

- **DS:** 235,000 sentences annotated with Distant Supervision from Freebase relations(7)
- **DS + CT:** the DS corpus with the 2,050 Crowd dev sentences added to it
- **CTLP:** the *CrowdTruth label propagation* dataset, with relation scores propagated over the DS data with the *label update function DS** using the 2,050 Crowd dev sentences

(7) Riedel et al. “Relation extraction with matrix factorization and universal schemas.” NAACL. 2013

Evaluation Results





CrowdTruth.org // github.com/CrowdTruth/Open-Domain-Relation-Extraction // data.CrowdTruth.org

VU UNIVERSITY AMSTERDAM

CrowdTruth The framework for crowdsourcing ground truth data

Google