Knowledge Extraction and Inference from Text: Shallow, Deep, and Everything in Between[1]

(Tutorial at SIGKDD 2018)

Part II

Partha Talukdar
IISc Bangalore and Kenome
ppt@iisc.ac.in

Tutorial homepage: https://goo.gl/vRkwxZ

Last updated: August 19, 2018

[1] Many slides reused from CIKM 2017 tutorial with Soumen Chakrabarti

Outline

13:00-13:15 Overview and motivation

13:15-13:45 Case study: NELL

13:45-14:00 Bootstrapped Entity Extraction

14:00-15:00 Open Relation Extraction & Canonicalization

15:00-15:30 Coffee Break

15:30-16:15 Distantly-supervised Neural Relation Extraction

16:15-16:45 Knowledge Graph Embeddings

16:45-17:00 Conclusion & QA

Introduction: Relation Extraction

Definition: A relation is defined in the form of $r(e_1, e_2)$, where the e_i are entities in a predefined relation r. Easily extended to n-ary relations (events).

Examples

- isAcquiredBy relationship between pairs of companies, e.g., isAcquiredby(Google, YouTube)
- isAppointedCeoOf relationship between a person and company
- geneCausesDisease between gene and disease

Types

- $?(e_1, e_2)$
- $r(e_1, ?)$
- $r(?, ?)$
- Macro (corpus-level) vs Micro (sentence-level)

Example relations

- "is acquired by" relationship between pairs of companies
- "is appointed CEO of" relationship between a person and company,
- "is employee of" relationship between a person and an organization
- ACE Task
  - "located at"
  - "near",
  - "part",
  - "role",
  - "social"

  over pairs from five top-level entity types "person", "organization", "facility", "location", and, "geo-political entity".
- BioCreAtIvE II Protein-Protein Interaction
  - gene-disease relations,
  - protein-protein interaction, and

Dominant Approaches to Relationship Extraction

- Supervised
  - For each relation type, we collect annotated sentences as training examples.
  - Preferred approach if the types of relations of interest is a small set
  - Pros: Several effective methods: CRF, LSTM, Bi-LSTM etc. Works quite well when enough training data is available.
  - Cons: Human effort required scales with the number of distinct relation types. Not feasible at scale.
- Weakly-supervised
  - Supervision is provided at the relation instance level, not annotated sentences (next slide).
  - Pros: Supervision size is small and easy to provide. Much more scalable and practical.
  - Cons: There is more noise in the resulting training data, results in a more challenging learning problem.
  - State-of-the-art is at around 60% Precision at 30% recall $\rightarrow$ lot of headroom for improvement

(Weak) supervision setup

Given a corpus D, set of relationship types $r_1, \ldots r_k$, entity types T_{r1}, T_{r2} forming arguments of relationship type r, and a seed set S of examples of the form (E_{i1}, E_{i2}, r_i) $1 \leq i \leq N$ indicating that E_{i1} has relationship r_i with E_{i2}.

E_1	E_2	r	Label
Alon Halevy	Anhai Doan	IsPhDAdvisorOf	+
Donald Knuth	Andrei Broder	IsPhDAdvisorOf	+
Jeff Ullman	Surajit Chaudhari	IsPhDAdvisorOf	+
Alon Halevy	Dan Suciu	IsPhDAdvisorOf	-
Google	YouTube	Acquired	+
Google	Yahoo!	Acquired	-
Microsoft	Powerset	Acquired	+

Sources of relation extraction features

- Surface tokens and their shapes
- Part of speech and chunk tags
- Constituency and dependency parses
- Word, type, relation embeddings (later)

Surface Tokens

The tokens around and in-between the two entities often hold strong clues for relationship extraction.

> ⟨*Company*⟩ *Kosmix* ⟨*/Company*⟩ *is located in the* ⟨*Location*⟩ *Bay Area* ⟨*/Location*⟩

A "is_situated" relationship between a Company entity and Location entity indicated by token "located" and bigram tokens "located in"

> *. . . the Center for Disease Control and Prevention, which is in the front line of the world's response to the deadly* ⟨*Disease*⟩ *Ebola* ⟨*/Disease*⟩ *epidemic in* ⟨*Location*⟩ *Zaire* ⟨*/Location*⟩*,*

A "outbreak" relationship between a disease and location is indicated by keyword "epidemic".

Part of speech tags

Verbs in a sentence are key to defining the relationship between entities, that are typically nouns or noun phrases.

⟨Location⟩ The University of Helsinki ⟨/Location⟩ hosts ⟨Conference⟩ ICML ⟨/Conference⟩ this year.

Word "hosts" as a verb is a clue..

The/DT University/NNP of/IN Helsinki/NNP hosts/VBZ ICML/NNP this/DT year/NN

Syntactic parse tree structure

⟨Location⟩ Haifa ⟨/Location⟩, located 53 miles from ⟨Location⟩ Tel Aviv ⟨/Location⟩ will host ⟨Conference⟩ ICML ⟨/Conference⟩ in 2010.

This tree brings "ICML" closer to "Haifa" than "Tel Aviv" because "Haifa" is the head of the noun phrase "Haifa, located 53 miles from Tel Aviv" which forms the subject of the verb phrase "will host ICML in 2010".

Syntactic parse tree structure

```
(ROOT
  (S
    (NP
      (NP (NNP Haifa))
      (VP (VBN located)
        (PP
          (NP (CD 53) (NNS miles))
          (IN from)
          (NP (NNP Tel) (NNP Aviv)))))
    (VP (MD will)
      (VP (VB host)
        (NP
          (NP (NNP ICML))
          (PP (IN in)
            (NP (CD 2010)))))))))
```

Dependency graph

Links each word to the words that depend on it

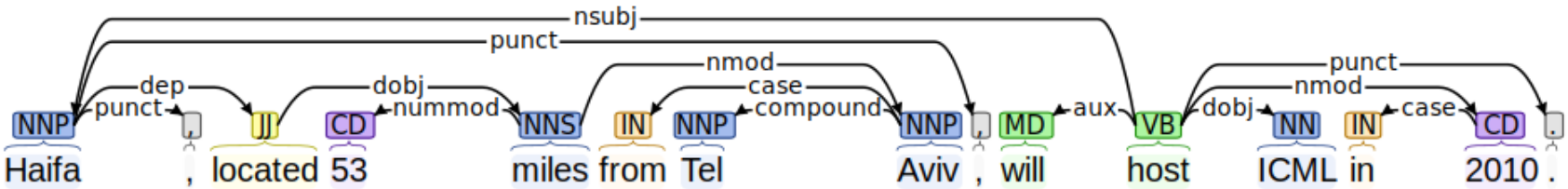


Figure : Dependency parse of a sentence.

- Verb “host” is linked to by both “Haifa” a location entity and with “ICML” a conference entity and this directly establishes a close connection between them
- In contrast, the path between ICML and Tel Aviv goes through “Haifa” and “Located”.

Weak Supervision Method Type 1: Bootstrapping

- Start with a small table of known facts

Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

- Find mentions of known authors and books in the corpus:

  The Robots of Dawn is a "whodunit" science fiction novel by Isaac Asimov, first published in 1983. It is part of Asimov's Robot series.

- Induce and evaluate patterns on known data
  - `*prefix, author, middle, title, suffix*`
  - `<LI><B>title</B> by author (`
  - `<i>title</i> by author (`
- Find matches to patterns over corpus (scan/index?)
- Import confident extractions into database

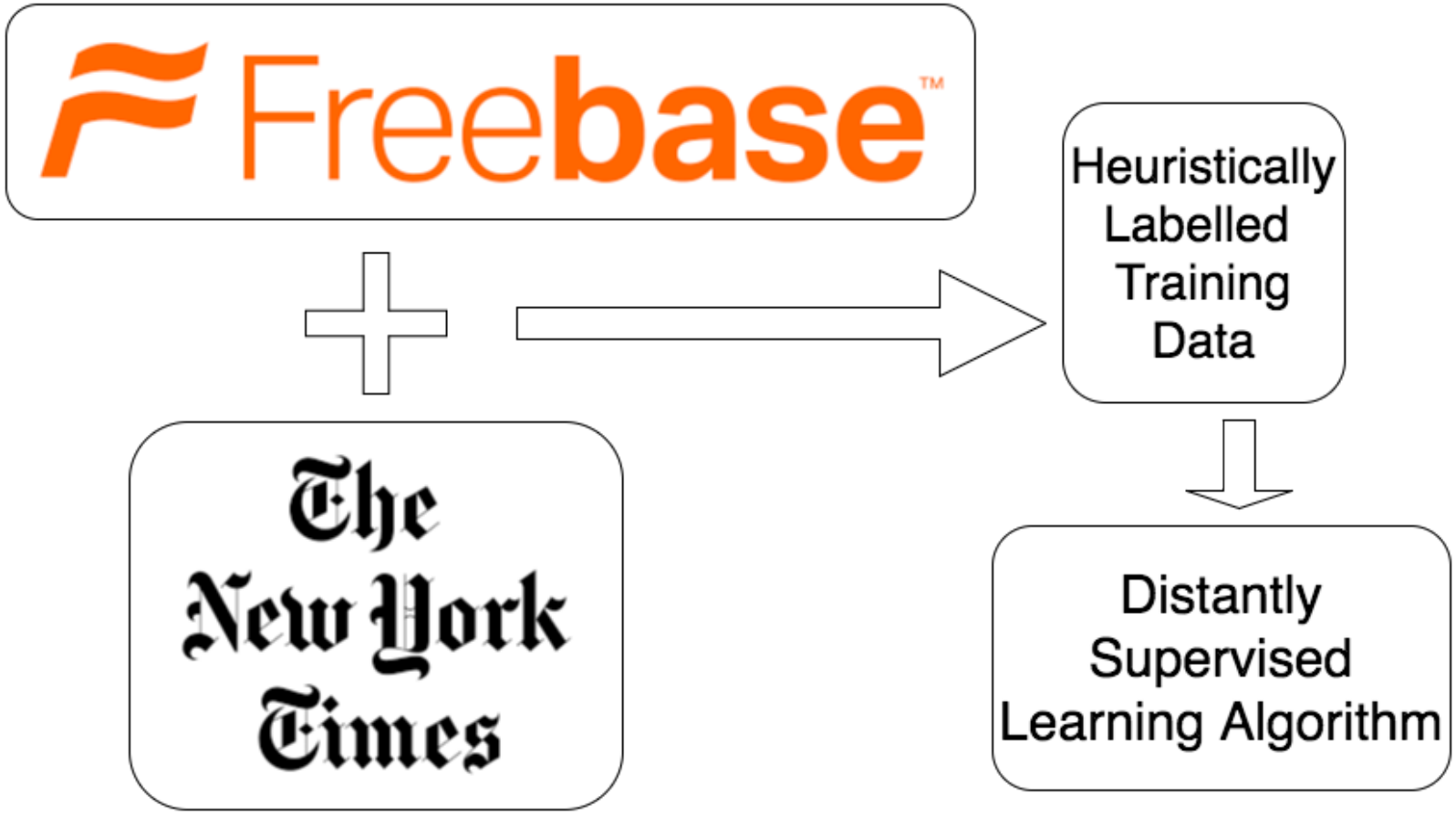
Weak Supervision Method Type 1: Bootstrapping (2)

- Rinse and repeat

1: input seed tuples $\{(e_{i1}, e_{i2}), i = 1, \ldots, n\}$
2: **while** database not big enough **do**
3: find snippets in corpus where seed tuples are mentioned
4: tag entities in snippets
5: generate new patterns L, t_1, C, t_2, R or L, t_2, C, t_1, R
6: apply new patterns over whole corpus
7: import newly extracted tuples into database

- Brin bootstrapped as follows: 5 facts $\rightarrow$ 199 occurrences $\rightarrow$ 3 patterns $\rightarrow$ 4047 proposed facts $\rightarrow$ 105 more patterns $\rightarrow$ 9369 proposed facts
- Quality control needed
  - Which extracted tuples are likely to be correct?
  - Which patterns are sufficiently reliable and useful?

Weak Supervision Method Type 2: Distant Supervision



Relation Extraction using Distant Supervision Example

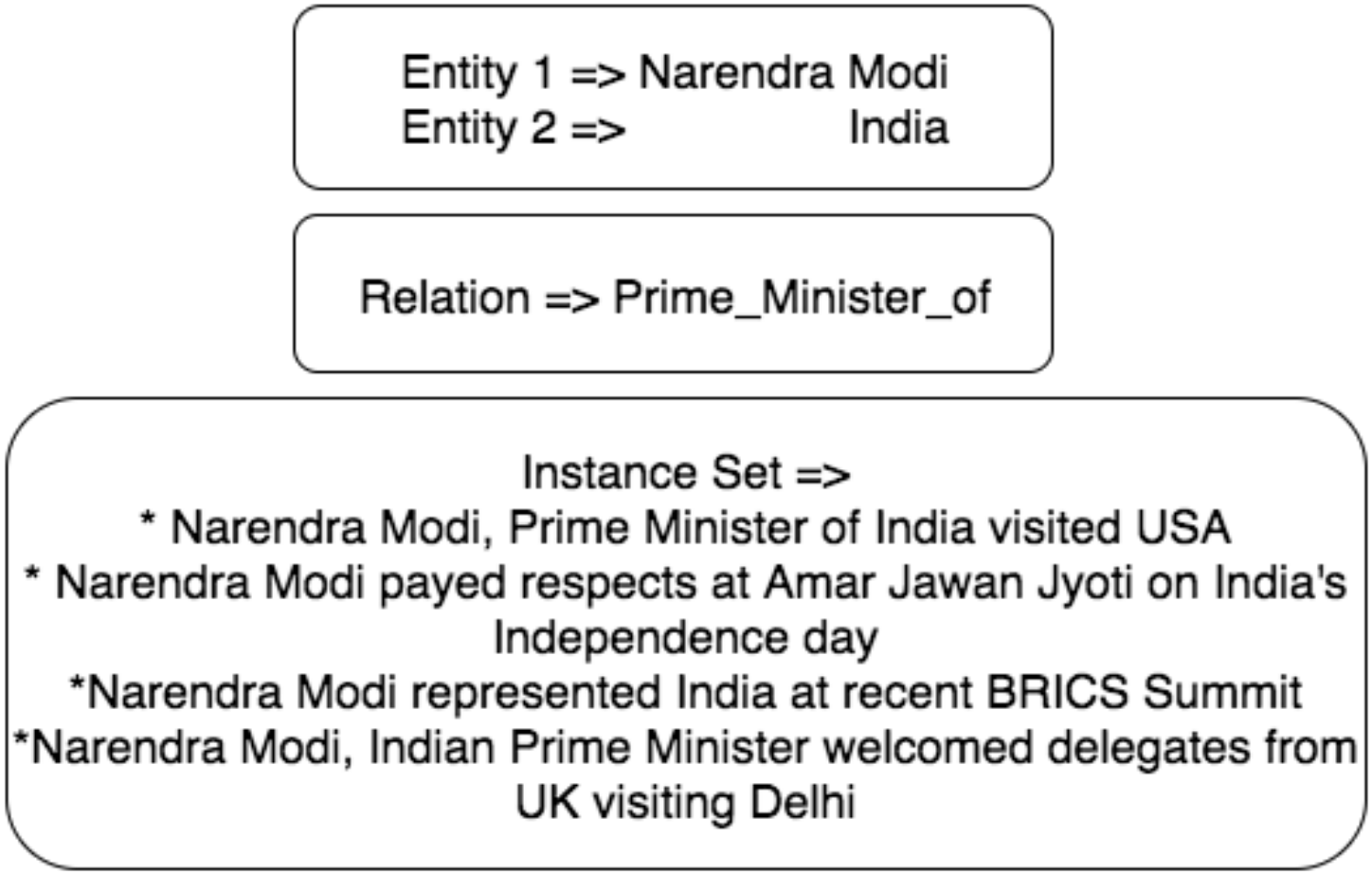


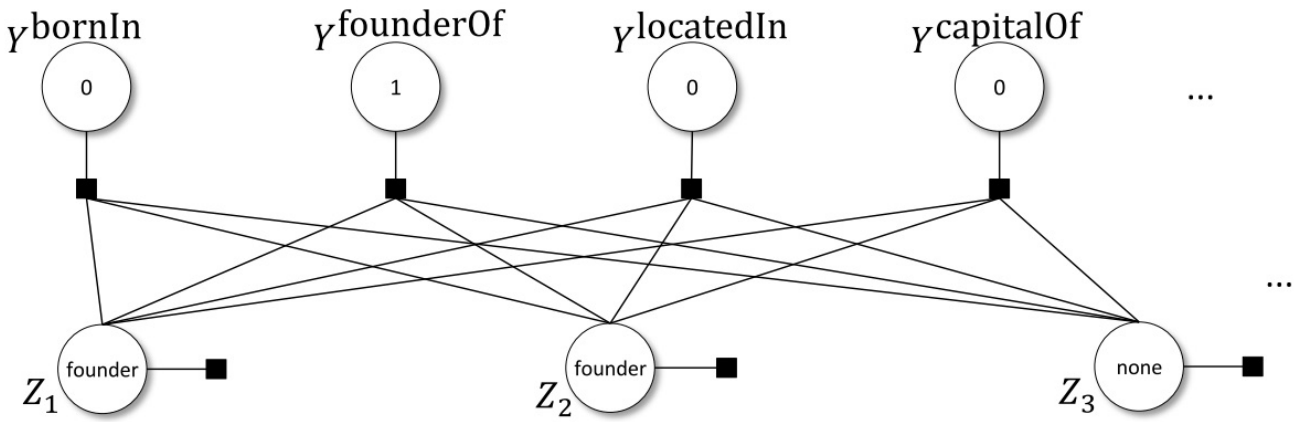
Figure : Example of an Instance Set in Distant Supervision training set

- Disadvantage: Noisy labelled data
- Advantage: Large amount of labelled data at very low cost

Multi-instance multi-label relation extraction

- [Mintz et al., 2009]: Distant Supervision (DS) as Supervised Learning. First work on DS.
- [Riedel et al., 2010]: Multi-instance (MI) formulation of DS
- [Surdeanu et al., 2012a]: Multi-instance Multi-label (MIML) DS formulation
  - Multiple relations may hold between entities e_1, e_2, evidenced in different sentences
  - If a relation r holds between entities e_1, e_2, then at least one sentence has to support with evidence
  - If relation r does not hold between entities e_1, e_2, then there can be no evidence sentence
  - Assuming all sentences are evidence of *some* relation pollutes training data

MultiR [Hoffmann et al., 2011]



Steve Jobs was founder of **Apple**.

Steve Jobs, Steve Wozniak and Ronald Wayne founded **Apple**.

Steve Jobs is CEO of **Apple**.

...

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}; \theta) \propto \prod_m \phi^{\text{extract}}(z_m, x_m; \theta) \prod_r \phi_r^{\text{join}}(y_r, \boldsymbol{z})$$

$$\phi_r^{\text{join}}(1, \boldsymbol{z}) = \bigvee\nolimits_m [\![z_m = r]\!]$$

$$\phi_r^{\text{join}}(0, \boldsymbol{z}) = \bigwedge\nolimits_r [\![z_m \neq r]\!]$$

$$\phi^{\text{extract}}(z_m, x_m; \theta) = \exp\left(\theta \cdot f(z_m, x_m)\right)$$

Neural Networks for Distant Supervision

- Prior to Neural Networks, all techniques utilised carefully curated NLP features for relation prediction.
- Extraction of traditional NLP features may creep in additional errors into the pipeline.
- Sentences encountered in relation extraction problem are on an average more than 40 words, which might lead to higher errors in NLP feature extraction [Zeng et al., 2015].
- To avoid errors in feature extraction from NLP pipeline, most modern systems use neural networks for extraction the features.

Piecewise Convolution Neural Network (PCNN)

- PCNN adapts CNN-based relation extraction [Zeng et al., 2014] to the distant supervision setting.
- CNNs are used to extract sentence features.
- Sentence features are then processed by a novel piecewise pooling method to preserve structural features of the sentence.
- Final features are then processed using a linear layer followed by softmax to generate probability of a given relation for an entity pair.

Piecewise Convolution Neural Network (PCNN) (2)

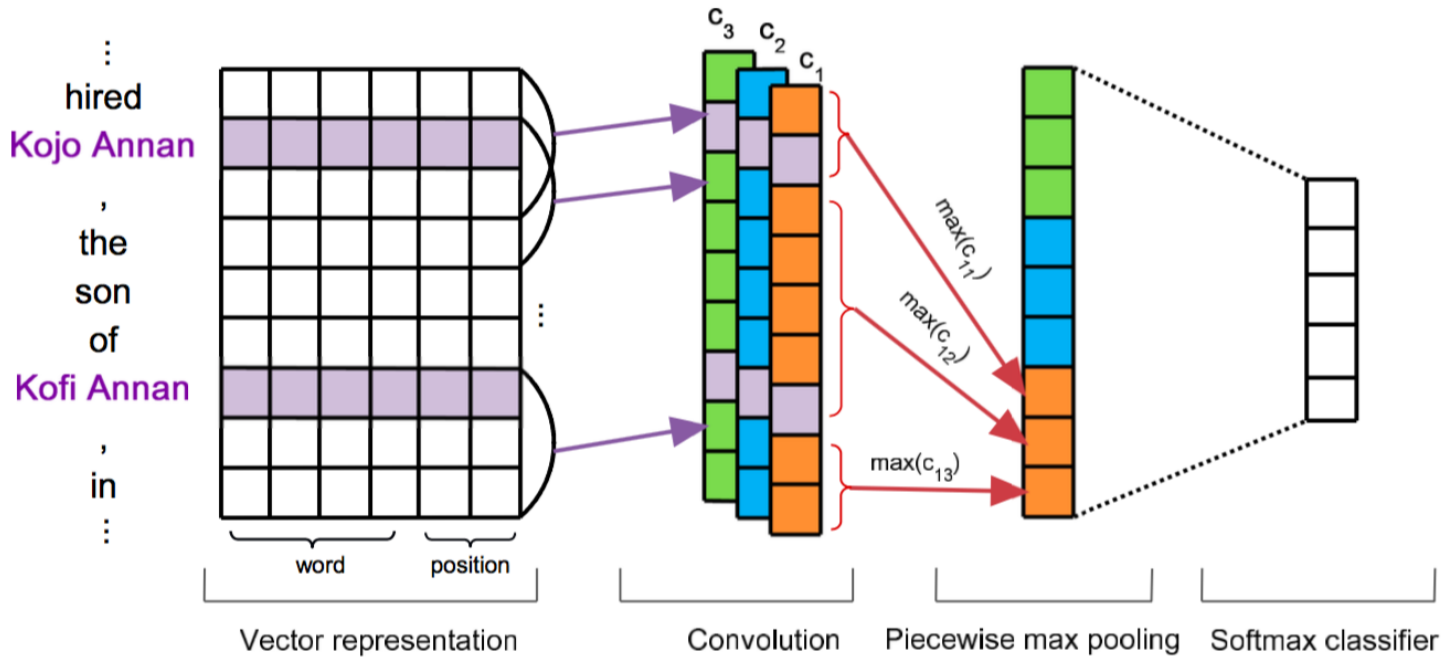


Figure : The architecture of PCNNs used for distant supervised relation extraction, illustrating the procedure for handling one instance of a bag and predicting the relation between Kojo Annan and Kofi Annan (Image from original paper)

Piecewise Convolution Neural Network (PCNN) (3)

- Let $Q_{i:j} \in \mathbb{R}^{(j-i+1)\times d}$ be the concatenation of word vectors from q_i to q_j
- Each word vector q_i is concatenation of word2vec embedding and the position embedding of the word.
- Convolution is a dot product between a filter $\Omega \in \mathbb{R}^{n\times d}$ and each consecutive n-gram in a stacked sequence of word vectors Q
- Dot product $c_i \in \mathbb{R}$ is defined as:

$$c_i = \Omega \times Q_{i-n+1:i},\ 1 \leq i \leq k+n-1$$

- Resulting $C_{\Omega,Q} = \langle c_1, \ldots, c_{k+n-1} \rangle$ is the embedding for the sentence.
- Maxpooling helps NN to go from variable sized sentences to fixed sized representations.

Piecewise Convolution Neural Network (PCNN) (4)

- To preserve structural features in a sentence, piecewise maxpooling is performed.
- Maxpooled $p_{\Omega,Q} \in \mathbb{R}$ corresponding to filter w is obtained as follows.

$$p_{\Omega,Q} = \langle \max(s_{[o,e1]}), \max(s_{[e1,e2]}), \max(s_{[e2,-1]}) \rangle$$

where $s_{[i,j]}$ is $[c_i, c_{i+1} \ldots c_j]$ and $e1, e2$ are the locations of the entities in the sentence.

- Maxpooled sentence representation is further processed using non-linearity layer, linear layer, followed by softmax layer to predict the relation label. [Zeng et al., 2015]

PCNN Performance Comparison

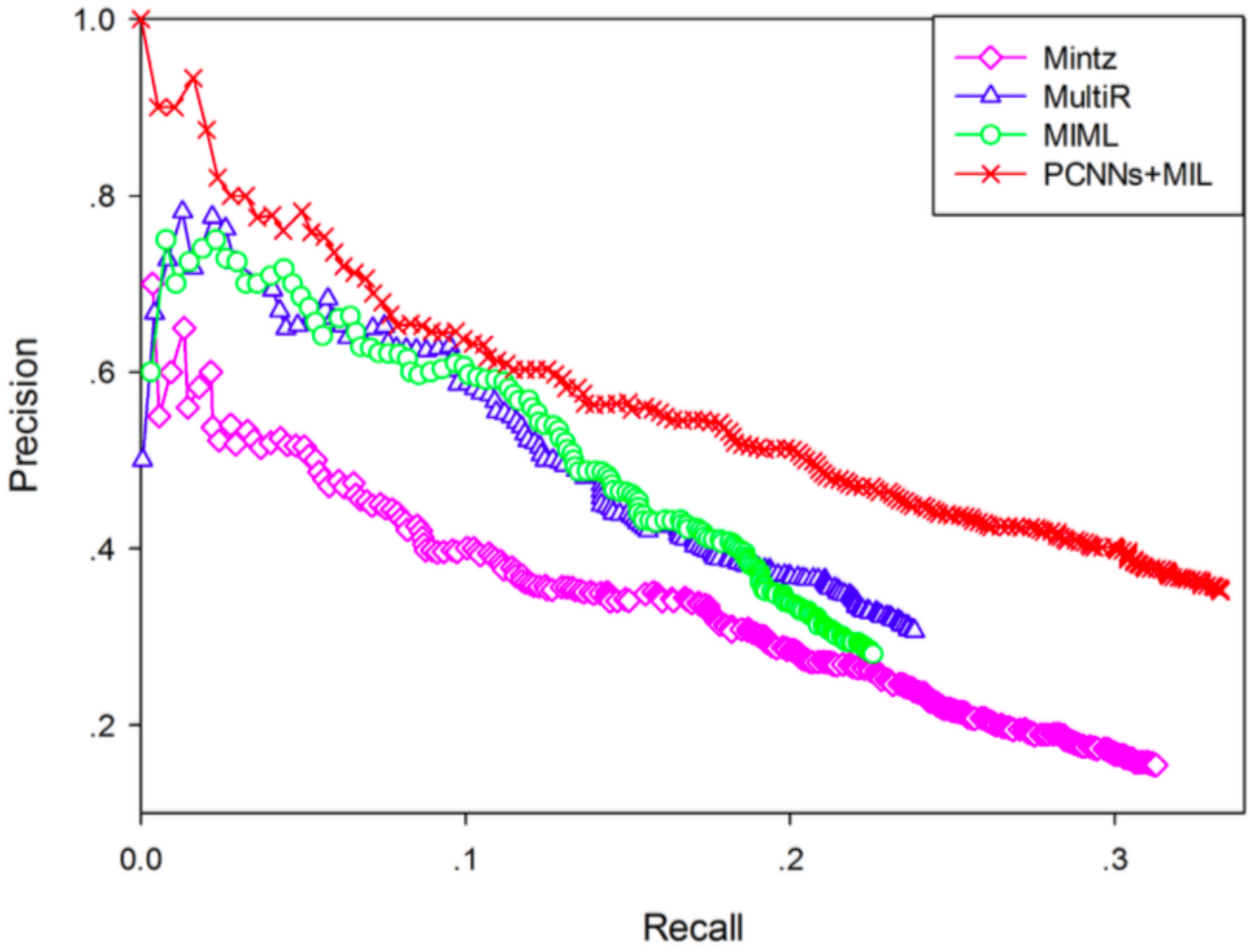


Figure : Performance comparison of the PCNN method with traditional approaches (Image from original paper)

Neural Relation Extraction (NRE) with Instance Attention [Lin et al., 2016a]

- PCNN model used a single sentence with highest relation probability to predict the label for an instance-set.
- Using single sentence from a bag to predict relation label, misses crucial information from other sentences for multi-relation prediction.
- NRE [Lin et al., 2016a] devises an attention mechanism to aggregate information from various sentence to form a single instance set representation.

Neural Relation Extraction (NRE) with Instance Attention [Lin et al., 2016a] (2)

- Suppose instance set $\mathbf{S}$ contains n sentences $x_1, x_2 \ldots x_n$.
- Representation for this instance set(s) is made as follows:

$$s = \sum \alpha_i x_i$$

- Attention values α is defined as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{1,k} \exp(e_k)}, e_i = x_i A r$$

  Where, A,r are parameters to be learned. A is a weighted diagonal matrix, and r is the query vector associated with relation r.
- Similar to PCNN, instance set representation (s) is then used as an input to a linear layer followed by softmax to predict multiple relations for each entity-pair.

Neural Relation Extraction with Selective Attention over Instances

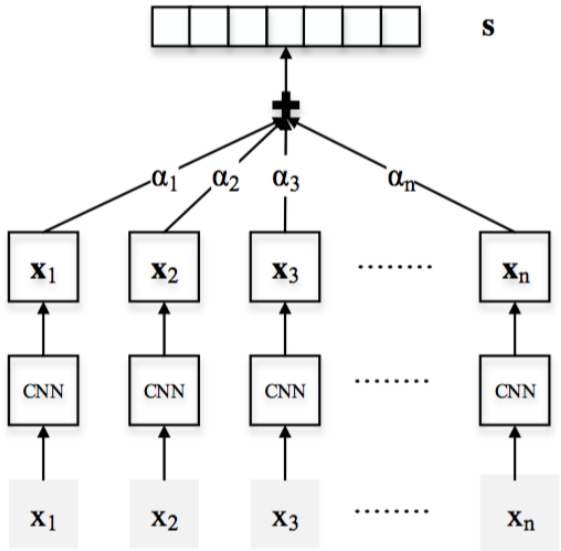


Figure : The architecture of sentence-level attention-based CNN, where x_i & $\mathbf{x_i}$ indicate the original sentence for an entity pair and its corresponding sentence representation, α_i is the weight given by sentence-level attention, and s indicates the representation of the sentence set (Image from original paper)

Neural Relation Extraction with Selective Attention over Instances

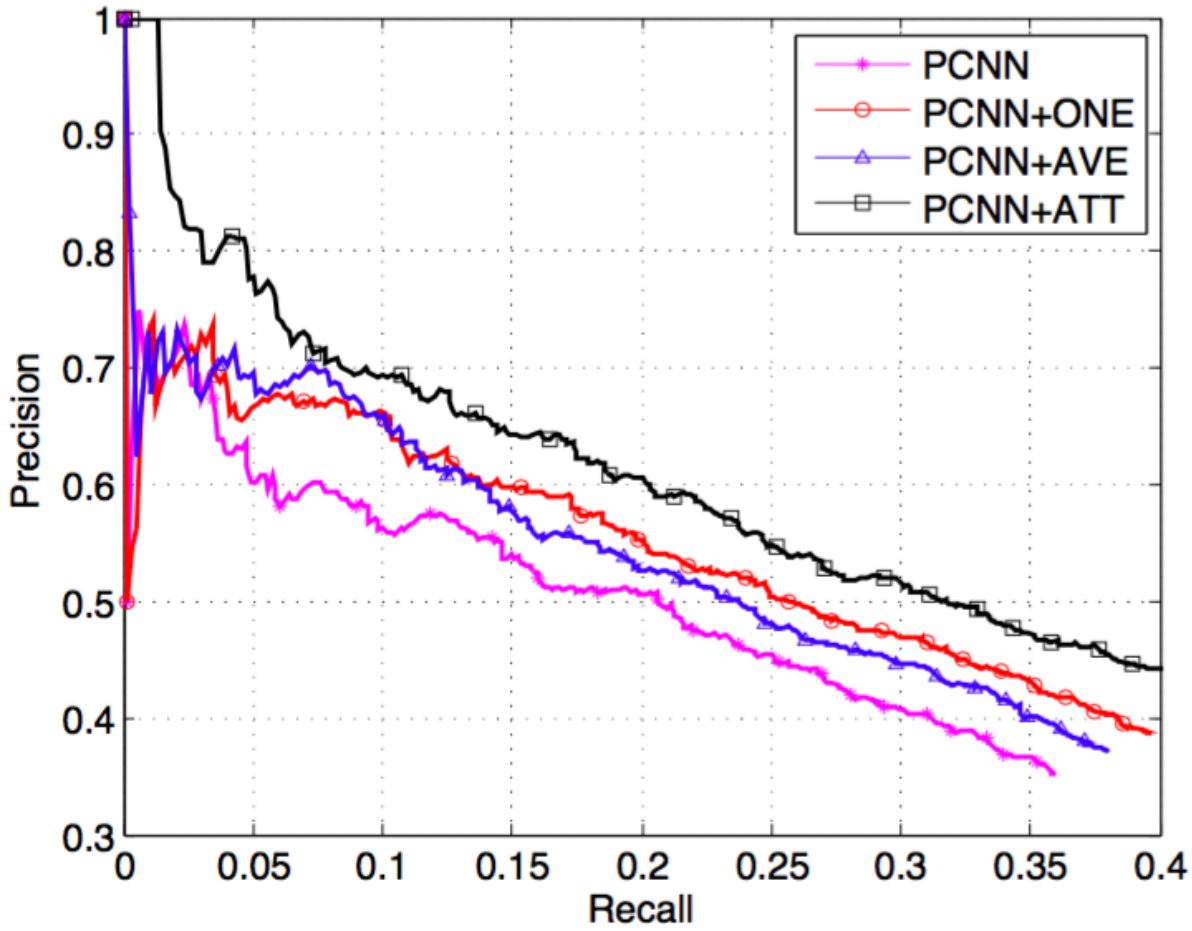


Figure : Performance comparison of the NRE method with traditional approaches (Image from original paper)

Bi-GRU based Word Attention Model (BGWA) [Jat et al., 2018]

- More than 50% of the sentences in Riedel Dataset contain greater than 40 words.
- Not all the words in a sentence express a given relation.
- Bi-GRU representation is combined using word attention.
- Assume u_{ij} to be the degree of relevance of the j^{th} word in i^{th} sentence of the instance set S as follows.

$$u_{ij} = w_{ij} \times A \times r;$$

$$a_{ij} = \frac{\exp(u_{ij})}{\sum_{l=1}^{M} \exp(u_{il})}$$

$$\hat{w_{ij}} = a_{ij} \times w_{ij}$$

Bi-GRU based Word Attention Model (BGWA) [Jat et al., 2018] (2)

- Bilinear operator A determines the relevance of a word for a relation vector r.

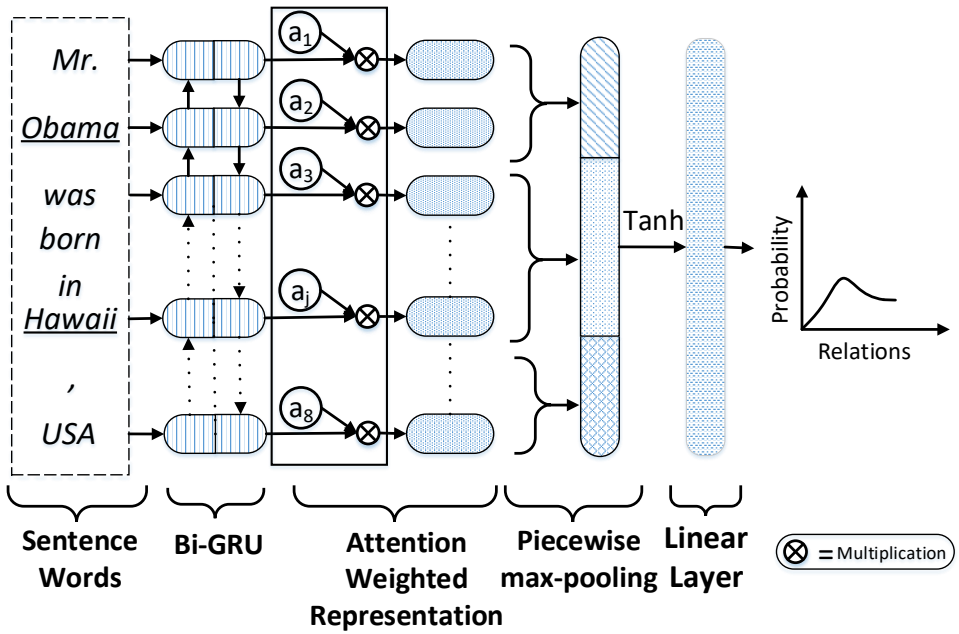


Figure : Bi-GRU word attention (BGWA) model(Image from original paper)

Entity based Word Attention Model (EA) [Jat et al., 2018]

- Information about entity can help the relation extractor.
- Entity-specific attention is applied by concatenating the entity embedding to each word and applying bi-linear attention.
- $[x_{ij}, e_k^{emb}]$ is the concatenation of a word and the entity embedding

$$u_{i,j,k} = [x_{ij}, e_k^{emb}] \times A_k \times r_k$$

$$a_{i,j,k} = \frac{\exp(u_{i,j,k})}{\sum_{l=1}^{M} \exp(u_{i,l,k})}$$

$$\hat{w}_{i,j,k} = a_{i,j,k} \times x_{i,j}$$

Entity based Word Attention Model (EA) [Jat et al., 2018] (2)

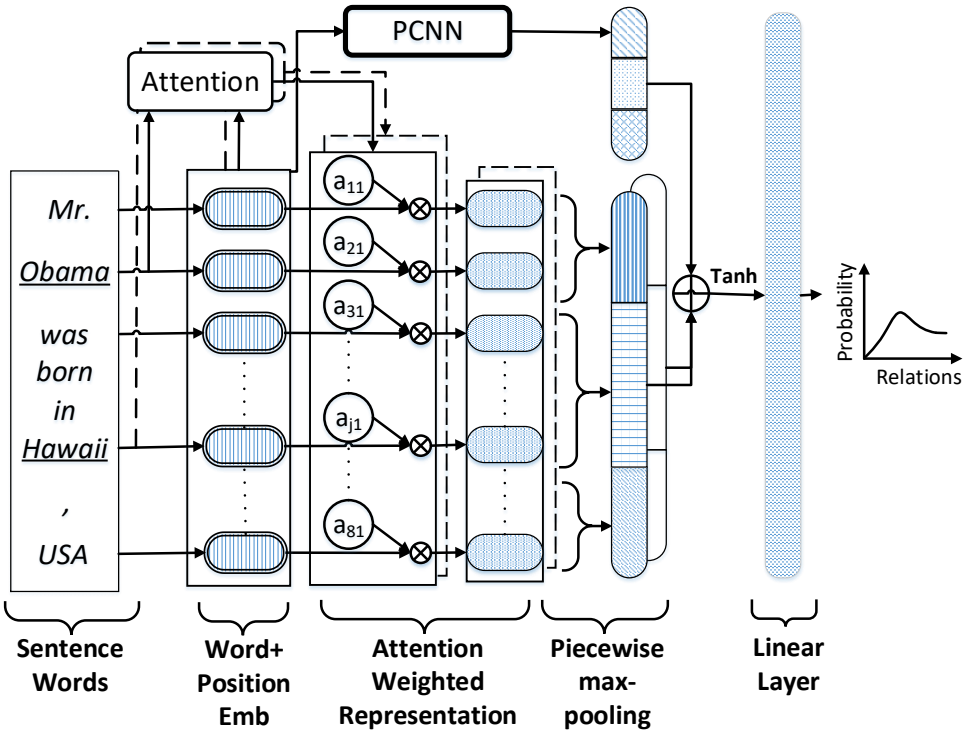


Figure : Entity based word attention (EA) model(Image from original paper)

Ensemble of Relation Extractors [Jat et al., 2018]

- PCNN, EA and BGWA models have complimentary strenghts.
- Combine the relation probabilities using simple linear model to achieve better results.

$$P_{i,\text{Ensemble}} = \alpha P_{i,\text{PCNN}} + \beta P_{i,\text{BGWA}} + \gamma P_{i,\text{EA}}$$

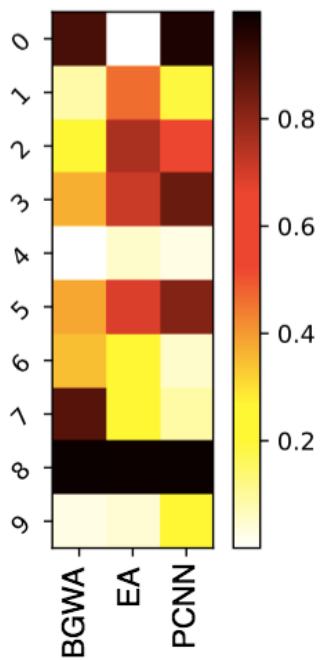


Figure : Confidence scores (indicated by color intensity, darker is better) of models on true labels of 10 randomly sampled instance sets from GIDS dataset. Rows represent the instance sets and columns represent the model used for prediction. The heatmap shows complementarity of these models in selecting the right relation. Motivated by this evidence, the proposed Ensemble method combines the three models, viz., Word Attention (BGWA), Entity Attention (EA) and PCNN (Image from original paper)

Ensemble of Relation Extractors [Jat et al., 2018] (2)

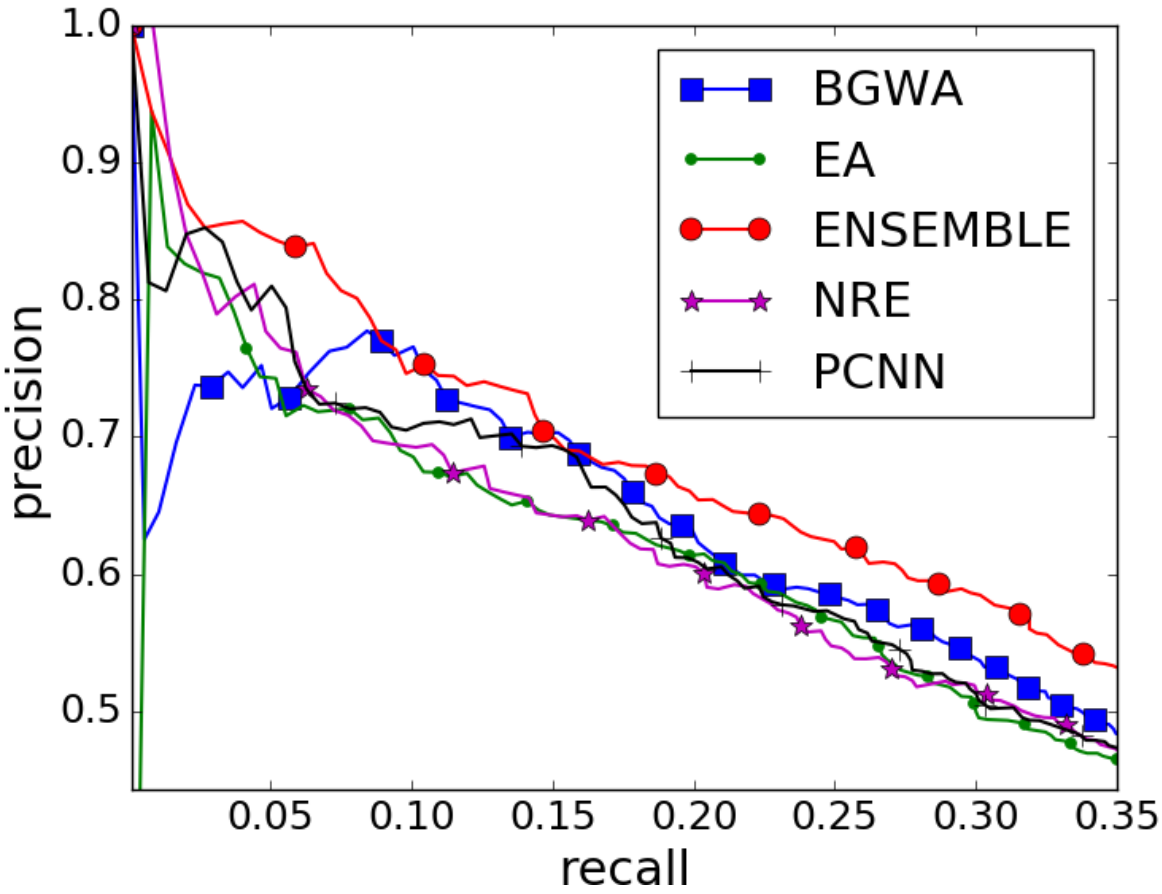


Figure : Model performance on Riedel 2010 dataset(Image from original paper)

Global Distant Supervision for Relation Extraction [Han, 2016]

- Indirect supervision using joint inference across relation instances by:
  - Consistency between relation and argument.
  - Consistency between neighbor instances.
  - Consistency between multiple relation labels.
- They use Markov Logic Networks to encode complex dependencies.
- Markov Logic Networks (MLN) [Richardson and Domingos, 2006]: probabilistic extension of first-order logic

Global Distant Supervision for Relation Extraction [Han, 2016] (2)

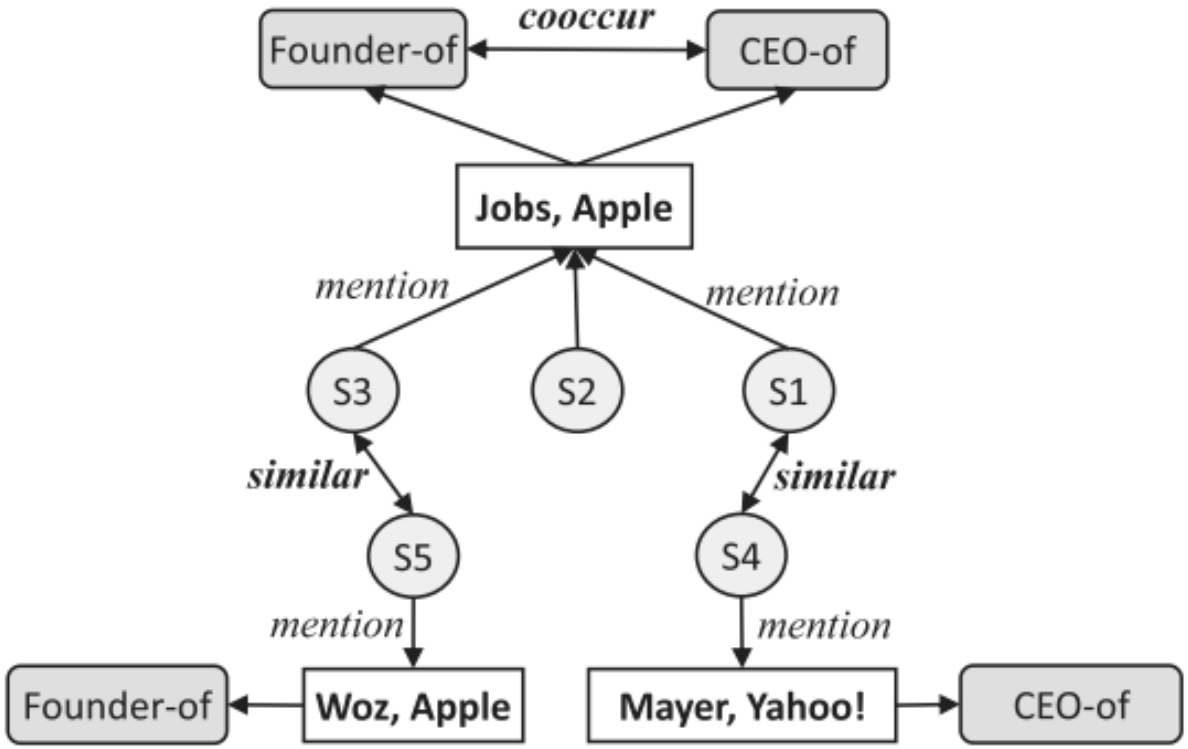


Figure : Dependencies between objects in Knowledge Base(Image from original paper)

Global Distant Supervision for Relation Extraction [Han, 2016] (3)

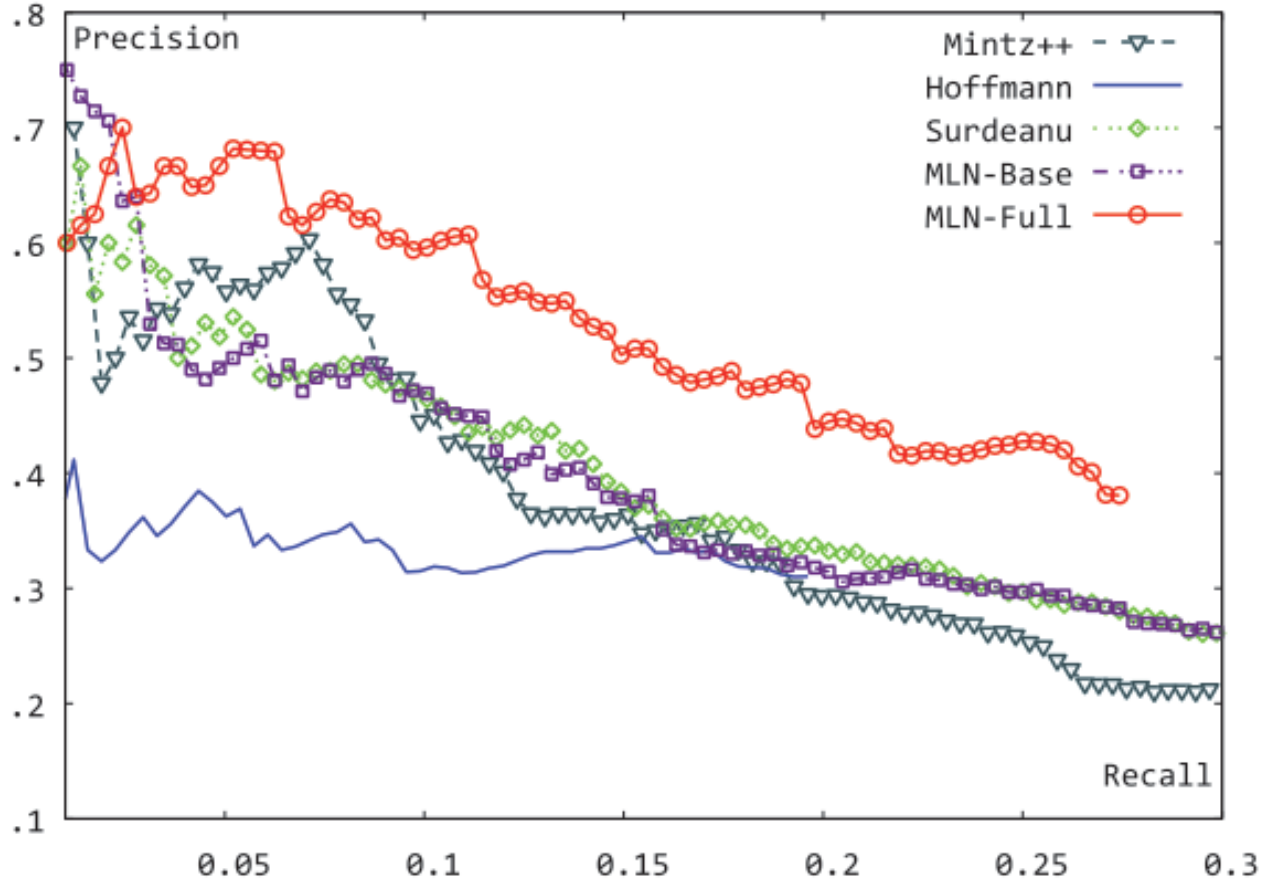


Figure : Precision-Recall curves on KBP dataset developed by [Surdeanu et al., 2012b].(Image from original paper)

Distant Supervision for Relation Extraction beyond the Sentence Boundary [Quirk and Poon, 2017]

- Previous distant supervision approaches are limited to SINGLE sentence proessing.
- Single sentence approach loses out on identifying some relations in the long knowledge tail.
- Document level graph representation with dependencies for adjacent sentences and discourse relations.
- Document Graph consisting of:
  - Nodes: words
  - Edges: Dependency, Adjacency, Discourse relations.
- Experiments were run on biomedical literature with distant supervision from Gene Drug Knowledge Database (GDKD).
- Cross-sentence extraction obtained far more unique relations compared to single-sentence extraction, improving absolute recall by 89-102%.

Noise Mitigation for Neural Entity Typing and Relation Extraction [Yaghoobzadeh et al., 2017]

- The paper handles two types of noise in Information Extraction.
  - Noise from Distant Supervision.
  - Noise from pipeline input features.
- The work integrates probablilistic output from entity type prediction into relation extraction.
- Probabilistic input helps the extraction system to compensate for errors during typing.
- Noise in relation extraction is mitigated using entity type probabilities.
- Experiments were performed on CF-Figment dataset, derived from ClueWeb data with FACC1 annotated Freebase entities.
- The area of JOINT-TRAIN under the PR curve is 0.66, as compared to baseline PCNN's 0.48 (CF-FIGMENT dataset).

Reinforcement Learning for Relation Classification from Noisy Data [Feng, 2018]

- Noise mitigation using Reinforcement Learning to select high quality sentences which may express a relation from a bag of sentences in distant supervision.
- Sentence level relation extraction, as opposed to bag level extraction done typically in DS algorithms.
- Manual Evaluation Results for 300 sentences: Proposed Method Macro F1= 0.42, Accuracy = 0.64. (baseline CNN+ATT, Macro F1= 0.29, Accuracy = 0.56)

Reinforcement Learning for Relation Classification from Noisy Data [Feng, 2018] (2)

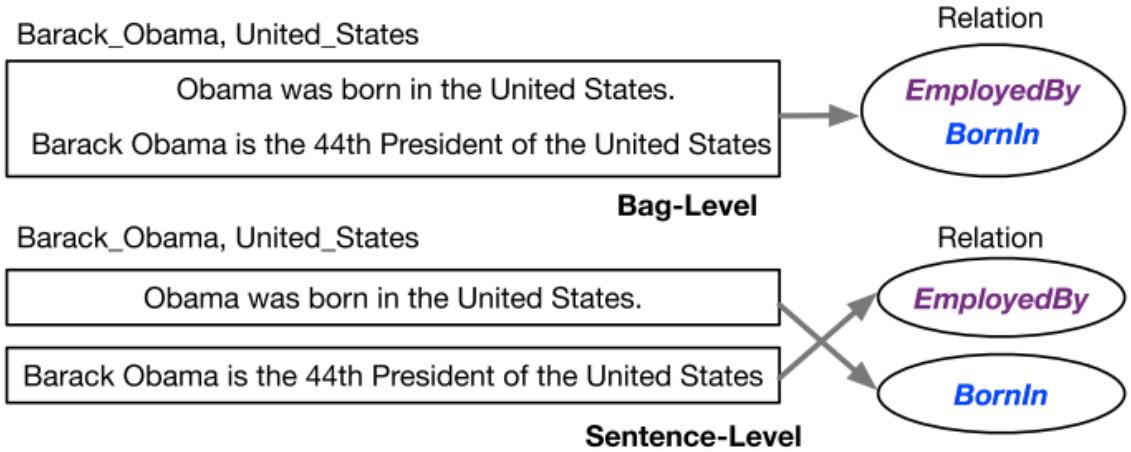


Figure : Sentence level Relation Extraction from Instance Bag.(Image from original paper)

- Instance selection problem: Given a set with pairs of (sentence, relation_label), relation r_i and entity-pairs (h_i, t_i)

$$X = (x_1, r_1), (x_2, r_2) \dots (x_n, r_n)$$

Reinforcement Learning for Relation Classification from Noisy Data [Feng, 2018] (3)

- Relation classification problem: Predict semantic relation r_i from sentence x_i. Model prediction $= P(r_i|x_i, h_i, t_i)$

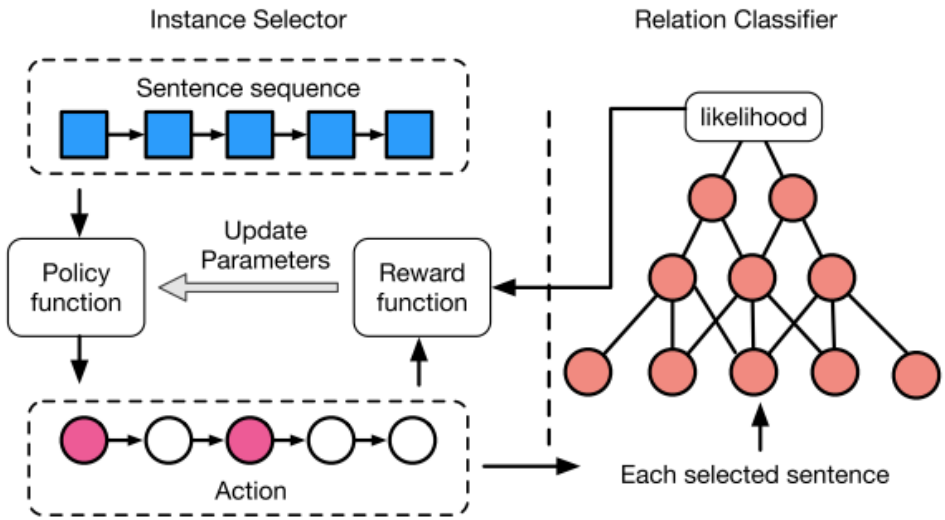


Figure : Sentence level Relation Extraction(Image from original paper)

DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction [Qin et al., 2018]

- **sentence-level noise reduction**: DSGAN learns a sentence level true-positive generator.
- Generator can be regarded as a special case of reinforcement learning based noise mitigation.
- A separate generator is trained for each relation.
- Experiments on the cleaned datasets are performed using baseline from [Lin et al., 2016b].
- DSGAN helps improve the Area Under the Curve (AUC) to 0.264 for PCNN+ATT model as compared to it's original value of 0.253.

DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction [Qin et al., 2018] (2)

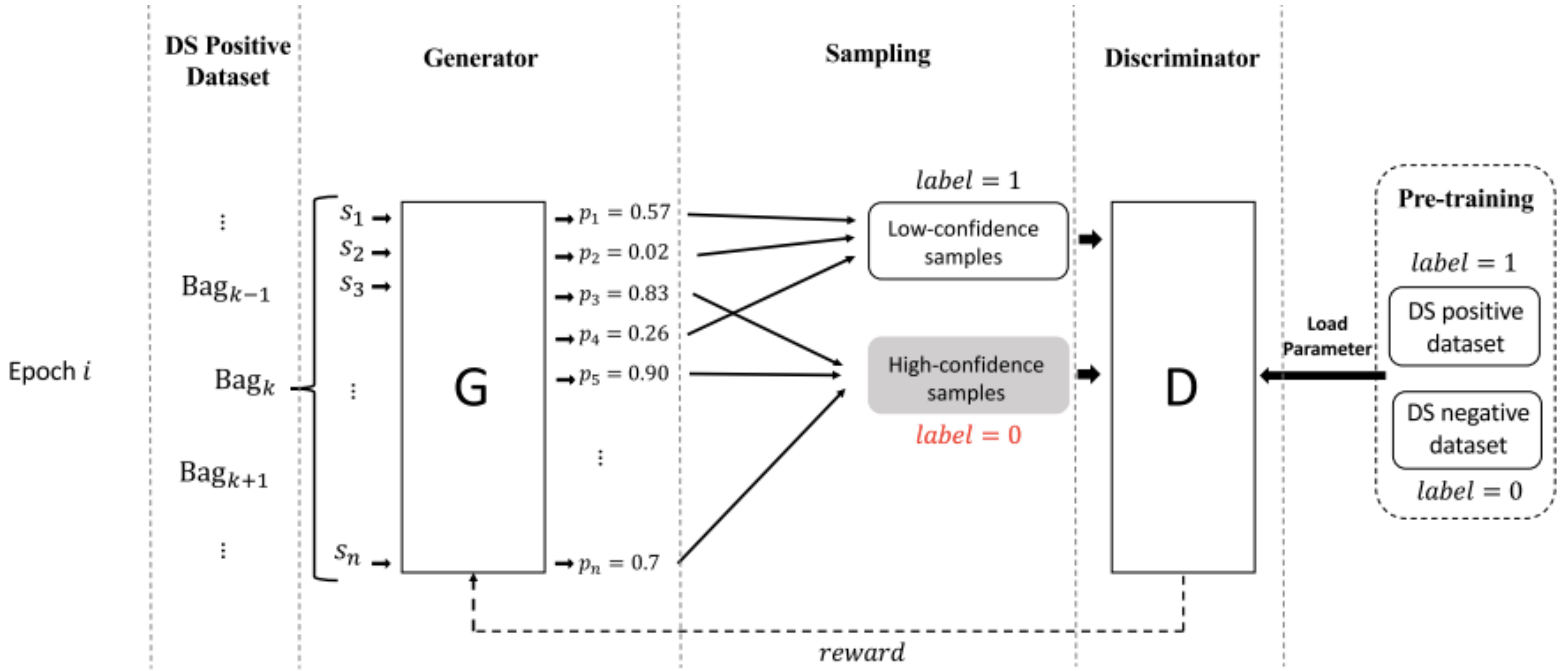


Figure 2: An overview of the DSGAN training pipeline. The generator (denoted by **G**) calculates the probability distribution over a bag of DS positive samples, and then samples according to this probability distribution. The high-confidence samples generated by G are regarded as true positive samples. The discriminator (denoted by **D**) receives these high-confidence samples but regards them as negative samples; conversely, the low-confidence samples are still treated as positive samples. For the generated samples, **G maximizes** the probability of being true positive; on the contrary, **D minimizes** this probability.

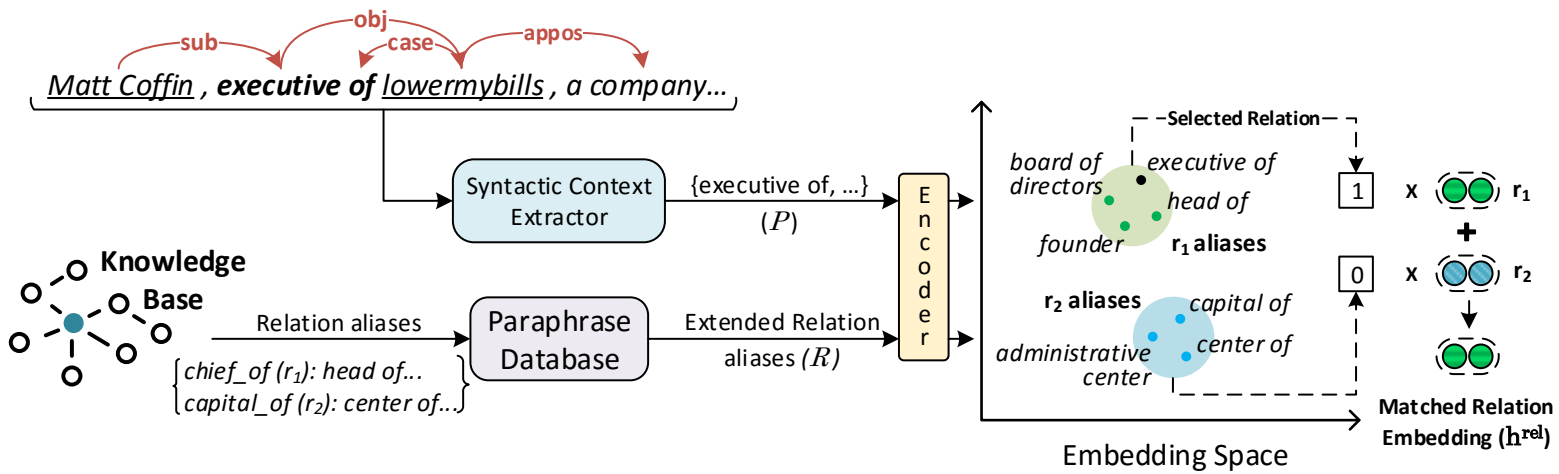
RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information [Vashishth et al., 2018b]

- Utilizes additional supervision from Knowledge Graph for improving distant supervised relation extraction.
- RESIDE uses Graph Convolution Networks (GCN) for modeling syntactic information and has been shown to perform competitively even with limited side information

RESIDE Overview:

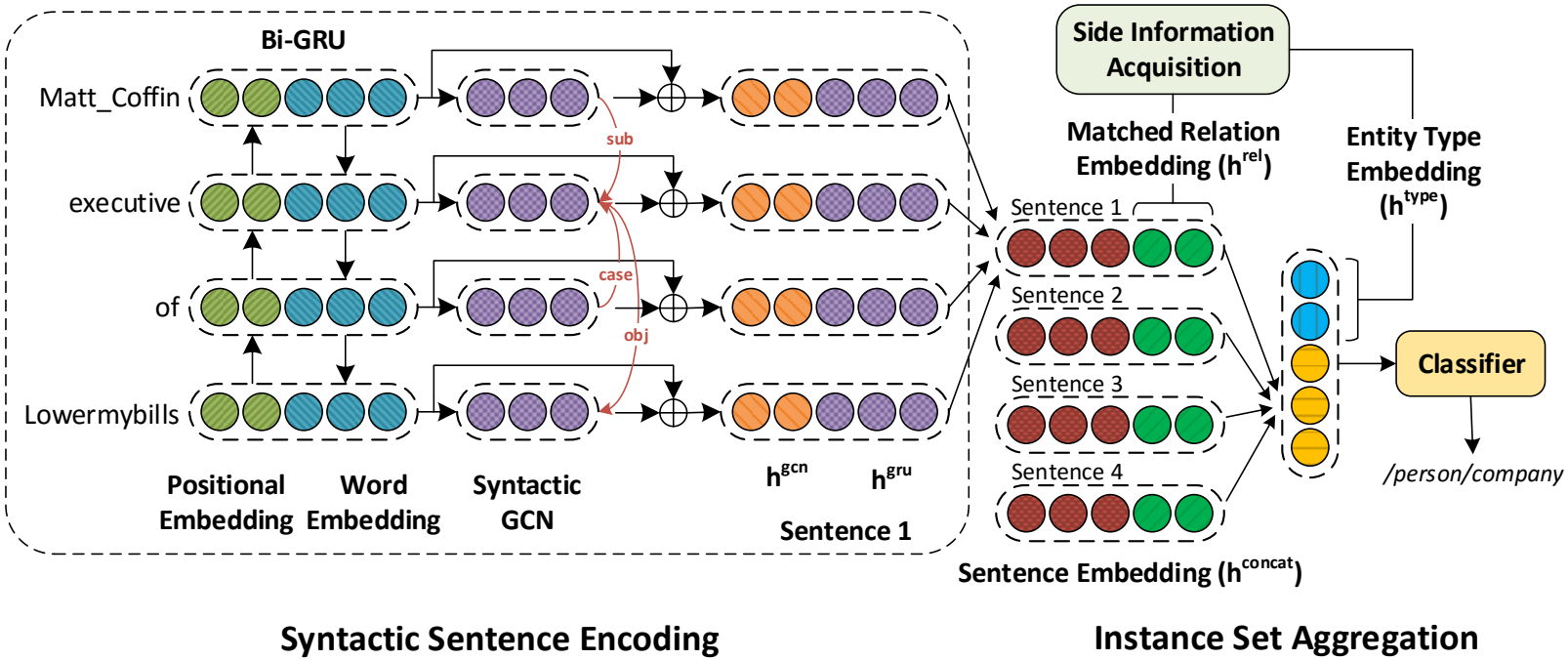
- **Syntactic Sentence Encoding:** Uses Bi-GRU and GCN for encoding each sentence.
- **Side Information Acquisition:** Utilizes additional supervision from KBs and Open IE methods for getting relevant side information.
- **Instance Set Aggregation:** Attention over sentences encoding to get a representation for the entire bag.

RESIDE: Side Information

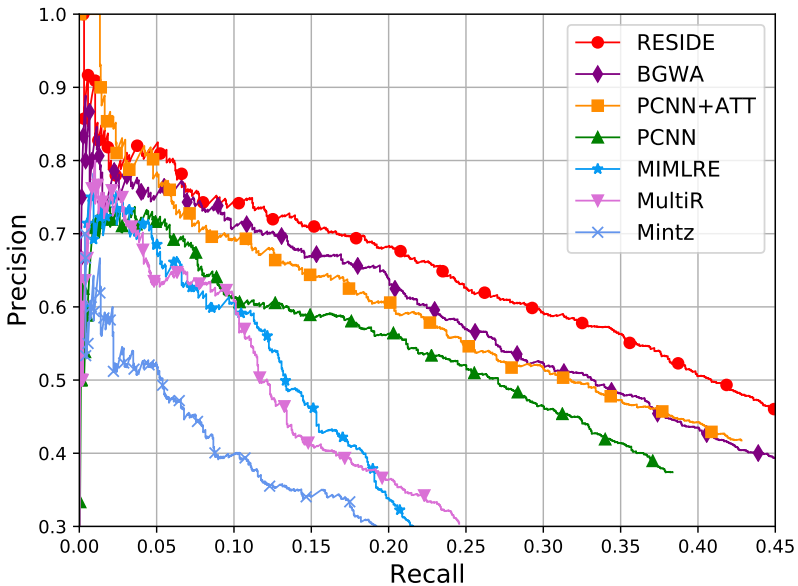


- **Relation Alias Side Information:** Extract relation phrases between target entities and links them to KG based on their closeness in embedding space.
- **Entity Type Side Information:** Utilizes entity type information for KG for putting constraints on predicted relation.

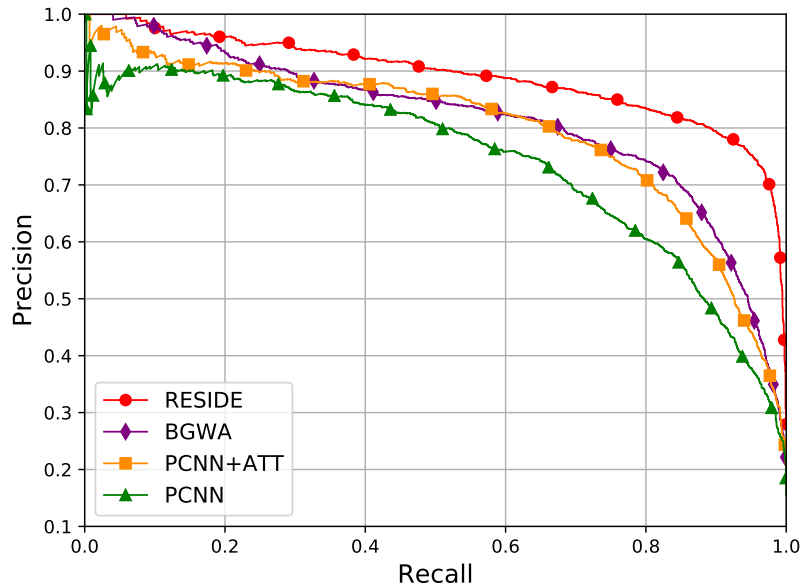
RESIDE: Improving Neural DS Relation Extraction using Side Information [Vashishth et al., 2018b]



RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information [Vashishth et al., 2018b]



(a) Riedel dataset



(b) GIDS dataset

Figure : Comparison of Precision-recall curve.

Datasets for Distantly Supervised Relation Extraction

Dataset	# relation	# sentences	# entity-pair
Reidel2010 Dataset [Riedel et al., 2010]			
Train	53	522,611	281270
Test	53	172,448	96,678
GIDS Dataset [Jat et al., 2018]			
Train	5	11297	6498
Dev	5	1864	1082
Test	5	5663	3247

Table : Statistics of various datasets available for Distantly Supervised Relation Extraction.

Resources

- Datasets:
  - NYT dataset [Riedel et al., 2010]: https://github.com/thunlp/NRE
  - Google-IISc (GIDS) dataset [Jat et al., 2018]: https://github.com/SharmisthaJat/RE-DS-Word-Attention-Models
- Source Code:
  - NRE [Lin et al., 2016b]: https://github.com/thunlp/NRE
  - Word Attention Models [Jat et al., 2018]: https://github.com/SharmisthaJat/RE-DS-Word-Attention-Models
  - RL for RE [Feng, 2018]: https://github.com/JuneFeng/RelationClassification-RL
  - Noise Mitigation [Yaghoobzadeh et al., 2017]: https://github.com/hayy2017/noise_mitigation
  - RESIDE: https://github.com/malllabiisc/RESIDE

Outline

13:00-13:15 Overview and motivation

13:15-13:45 Case study: NELL

13:45-14:00 Bootstrapped Entity Extraction

14:00-15:00 Open Relation Extraction & Canonicalization

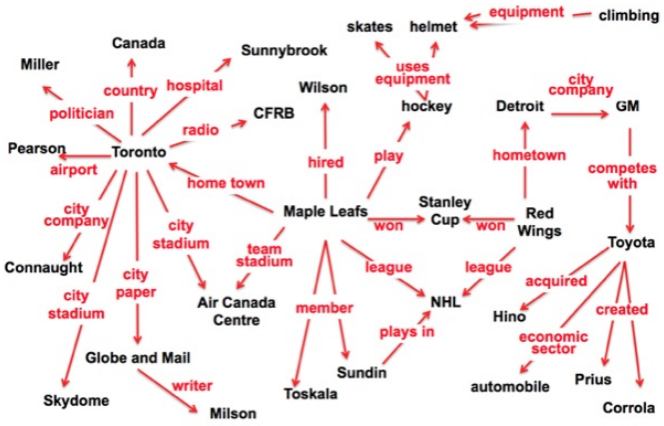
15:00-15:30 Coffee Break

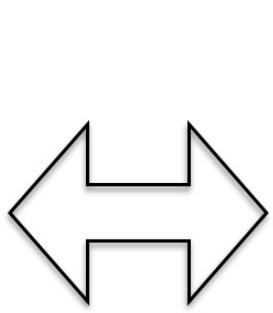
15:30-16:15 Distantly-supervised Neural Relation Extraction

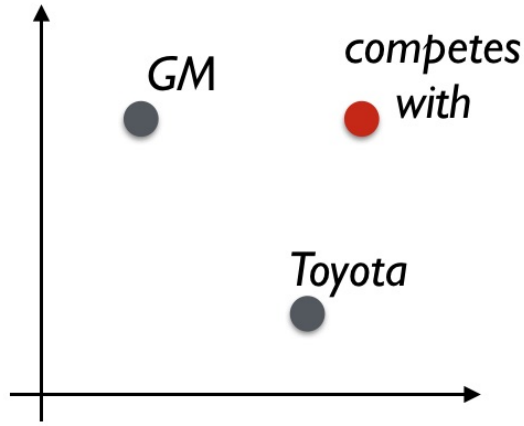
16:15-16:45 Knowledge Graph Embeddings

16:45-17:00 Conclusion & QA

Two Views of Knowledge







Knowledge Graph

Dense Representations

- The goal is to embed entities, types and relations, based on mostly-sound but largely-incomplete KGs fine types
- Embeddings can help many tasks, e.g., knowledge base completion (KBC), more accurate or extended fine type tagging or entity linking, inferring paraphrasing, entailment, or contradiction, etc.

Tasks and datasets for KG/KB embedding

WN18: 41k entities (WordNet synsets), 18 relation types (hypernymy, synonymy, ...), folds 141k/5k/5k

FB15k: 15k Freebase entities, 1345 relation types, folds 483k/50k/59k

FbSnapshots: Two snapshots of Freebase some time interval apart; removes sampling idiosyncrasies

NYT+FB: Freebase triples, plus dependency path-based textual relations from New York Times; entity mentions aligned with FB entity IDs; 25k entities, 4k relation types

FB15k+ClueWeb12: Corpus is ClueWeb12 with Google entity annotations

---

- WN18, FB15k, FbSnapshots used for knowledge base completion (KBC)
- Algorithm fits model using train and dev folds; ranks all other triples; those in eval fold are "relevant" docs; use ranking performance measures like MAP, MRR
- NYT+FB can be used for jointly embedding entities and relations informed by both KG and text

Structured embedding (SE)

- Entity can act as subject or object
- Accordingly each entity e gets two vectors $\vec{e}_s, \vec{e}_o$
- Each relation r is also associated with two matrices M_{rs}, M_{ro}
- If $M_{rs}\vec{e}_s$ and $M_{ro}\vec{e}_o$ are "close", then (e_s, r, e_o) is more likely, e.g.,

$$\Pr(e_s, r, e_o) = \sigma((M_{rs}\vec{e}_s) \cdot (M_{ro}\vec{e}_o))$$

- Or some decreasing function of $\|M_{rs}\vec{e}_s - M_{ro}\vec{e}_o\|$
- In words, each relation has two associated projections that bring participating entities close after projection
- Negative sampling: if (e_s, r, e_o) holds, replace with (uniformly?) randomly sampled (e'_s, r, e'_o) and assume these do not hold
- (Exactly one perturbed in each negative instance, not both)
- We want $\|M_{rs}\vec{e}_s - M_{ro}\vec{e}_o\| \ll \|M_{rs}\vec{e}'_s - M_{ro}\vec{e}'_o\|$

Structured embedding (SE) (2)

- Let $\gamma > 0$ be a margin hyperparameter
- Loss is defined as

$$\sum_{(e_s,e_o)} \sum_{(e'_s,e'_o)} \max\left\{0, \|M_{rs}\vec{e}_s - M_{ro}\vec{e}_o\| + \gamma - \|M_{rs}\vec{e}'_s - M_{ro}\vec{e}'_o\|\right\}$$

- What would happen if no margin were used ($\gamma = 0$)?
- Nonconvex; use gradient descent

TransE

- Each entity e associated with one vector $\vec{e}$
- Each relation r associated with one vector $\vec{r}$
- Model: if (e_s, r, e_o) holds, expect $\vec{e}_s + \vec{r} \approx \vec{e}_o$
- As with SE, perturb to e'_s, e'_o and expect $\|\vec{e}_s + \vec{r} - \vec{e}_o\| \ll \|\vec{e}'_s + \vec{r} - \vec{e}'_o\|$
- And therefore the loss to minimize is

$$\sum_{(e_s,e_o)} \sum_{(e'_s,e'_o)} \max\left\{0, \gamma + \|\vec{e}_s + \vec{r} - \vec{e}_o\| - \|\vec{e}'_s + \vec{r} - \vec{e}'_o\|\right\}$$

- Fewer parameters than SE
- Cannot model many-to-one, one-to-many, or many-to-many relations

TransR and STransE

- Mixes up projection and translation
- In TransR, a single projection matrix M_r is associated with each r, and applied to both $\vec{e}_s$ and $\vec{e}_o$
- Each relation r also associated with translation $\vec{v}_r$
- We expect $M_r\vec{e}_s + \vec{v}_r \approx M_r\vec{e}_o$ if (e_s, r, e_o) is valid in the KG
- STransE keeps M_{rs} distinct from M_{ro} as in SE, but retains the translation $\vec{v}_r$ of TransR
- I.e., we expect $M_{rs}\vec{e}_s + \vec{v}_r \approx M_{ro}\vec{e}_o$ if (e_s, r, e_o) is valid in the KG
- Loss and training similar to SE and TransE

TransH

- Each relation r associated with
  - A hyperplane defined by unit normal vector p_r
  - A translation vector d_r as in TransE
- Project e_s on to hyperplane ("$e_s \downarrow p_r$"), translate by d_r, compare with $e_o \downarrow p_r$
- Consider e_s, e_o, p_r with their tails at the origin
- $e_s \downarrow p_r = e_s - (e_s \cdot p_r)e_s$, likewise for e_o
- $\Delta(e_s, r, e_o) = \|(e_s \downarrow p_r) + d_r - (e_o \downarrow p_r)\|_2$
- Again, do pairwise training via perturbation: expect $\Delta(e_s, r, e_o) \ll \Delta(e'_s, r, e'_o)$

ITransF [Xie et al., 2017]

- STransE uses too many parameters per relation which is not good for rare relations
- ITransF allows parameter sharing among relations by using a set of underlying "concept" matrices stacked as tensor D
- Each relation r is associated with two attention vectors α_{rs} and α_{ro}
- $M_{rs} = \alpha_{rs} D$ and $M_{ro} = \alpha_{ro} D$
- We expect $M_{rs} e_s + r \approx M_{ro} e_o$ for valid triples in KG
- Loss and training similar to STransE

CP and Rescal decompositions

- Suppose e_s, e_o are column vectors
- Form outer product matrix $e_s e_o^\top \in \mathbb{R}^{D \times D}$
- Makes explicit feature crosses
- Relation r represented by $M_r \in \mathbb{R}^{D \times D}$
- Confidence that (e_s, r, e_o) holds in KG is large if $M_r \bullet (e_s e_o^\top)$ is large, and vice versa, where $\bullet$ is elementwise dot-product $\sum_{i,j} M_r[i,j] e_s[i] e_o[j] = e_s^\top M_r e_o$
- In general $e_s^\top M_r e_o \neq e_o^\top M_r e_s$, so asymmetry can be supported
- Some systems prefer a diagonal (therefore symmetric) matrix for M_r to reduce trainable weights . . . DistMult
- If we stack up the M_rs over all relations r, we get a 3-axis tensor $\boldsymbol{M} \in \mathbb{R}^{D \times D \times R}$
- If r, r' hold (and not hold) frequently between e_s, e_o, their "plates" in $\boldsymbol{M}$ should be similar

CP and Rescal decompositions (2)

- The 2-axis regular matrix analog would be, some rows (or columns) are similar to, or linearly dependent on, other rows (or columns)
- In which case, it would have low(er than full) rank
- Factorization via SVD reveals rank of a 2-axis matrix
- Defining and estimating rank of a tensor are more tricky
- Candecomp/Parafac is one approximate decomposition scheme:

$$\boldsymbol{X} \approx \sum_{i=1}^{I} \boldsymbol{a}_i \otimes \boldsymbol{b}_i \otimes \boldsymbol{c}_i$$

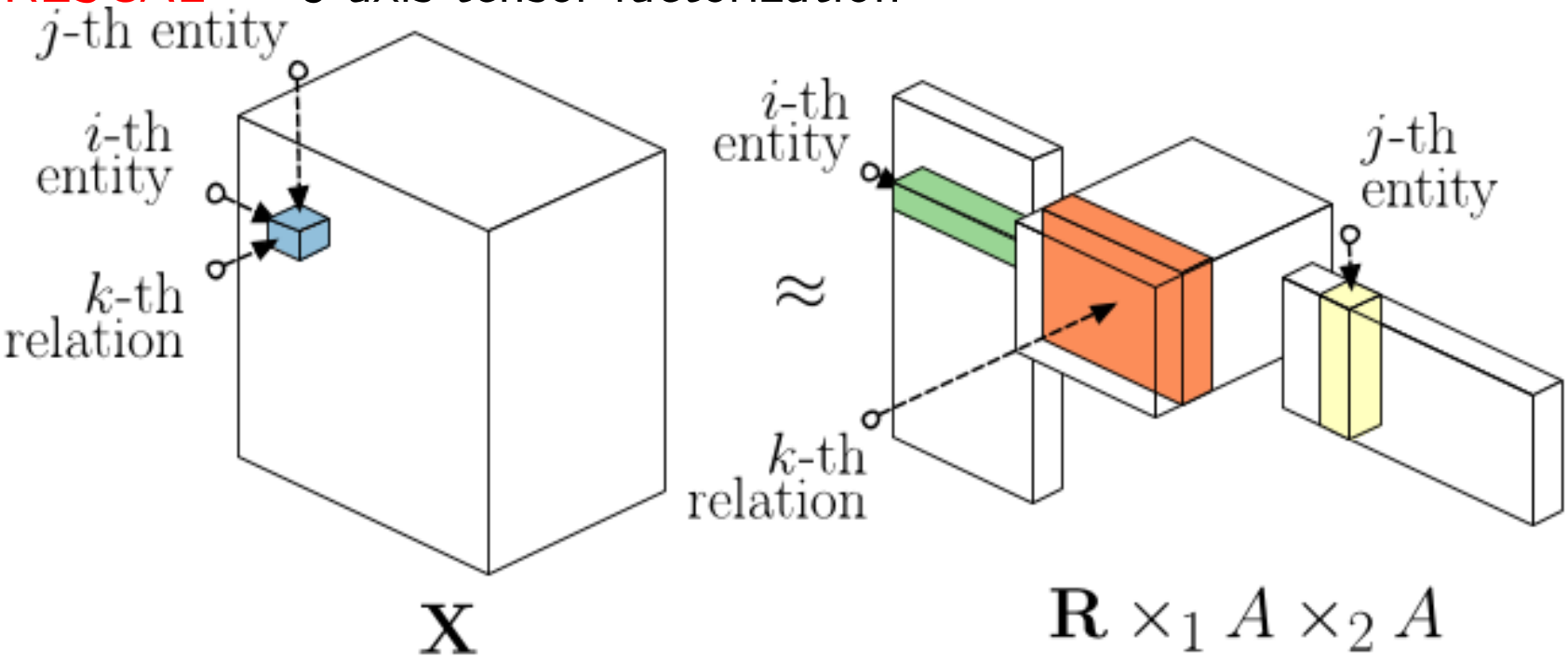
i.e.,

$$X[\ell, m, n] = \sum_{i} a_i[\ell] b_i[m] c_i[n]$$

where $\boldsymbol{X} \in \mathbb{R}^{L \times M \times N}$ and $\boldsymbol{a}_i \in \mathbb{R}^L, \boldsymbol{b}_i \in \mathbb{R}^M, \boldsymbol{c}_i \in \mathbb{R}^N$ and I is our control on the "rank" of approximating $\boldsymbol{X}$

CP and Rescal decompositions (3)

- RESCAL — 3-axis tensor factorization



- Assume N entities, R relations, D is the embedding dimension
- In tensor notation,

$$\boldsymbol{X} \approx \boldsymbol{M} \times_1 A \times_2 A$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times N \times R}, \boldsymbol{M} \in \mathbb{R}^{D \times D \times R}, A \in \mathbb{R}^{N \times D}$

CP and Rescal decompositions (4)

- For the rth plate,

$$X[r,:,:] \approx A^\top M[r,:,:]A$$

- Overall loss objective to minimize is

$$\underset{A,\boldsymbol{M}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{M} \times_1 A \times_2 A\|^2 + \lambda_A \|A\|^2 + \lambda_M \|\boldsymbol{M}\|^2$$

- Overall non-convex, use alternating least-squares
  - Fix A and improve $\boldsymbol{M}$
  - Fix $\boldsymbol{M}$ and improve A
- Can use SGD or batch solvers
- Quite a bit more compute-intensive than SVD/NMF style matrix decomposition

Asymmetry, antisymmetry, transitivity

- Entities are almost always single points/vectors
- Relations are translations (possibly after projection) ...
- ... or an inner product matrix (e.g., DistMult)
- One algebraic structure to fit all relation types
- But relations are diverse
  - Sibling-of is symmetric
  - Hypernym is asymmetric
  - Parent-of is antisymmetric
  - Subtype-of is transitive
  - Citizen-of is general many-to-few

ComplEx [Trouillon et al., 2016]

- Entities and relations are modeled as vectors in Complex domain
- Confidence of correctness of a triple is proportional to the complex dot product between e_s and e_o weighted by r

$$Pr(e_s, r, e_o) = \sigma(\Re(\langle r, e_s, \overline{e_o} \rangle))$$

- Similar to DistMult but in Complex domain
  - $\overline{e_o}$ is complex conjugate[2] of e_o
  - $\Re(\cdot)$ is the real part
- Allows handling symmetric, asymmetric and anti-symmetric relations together
- Logistic loss for training

[2] $\overline{a + ib} = a - ib$

HolE : Holographic Embedding [Nickel et al., 2016b]

- Entities and relations are modeled as vectors similar to DistMult
- Model is motivated by holographic models of associative memory and it learns compatibility between relations and entity pairs
- Confidence of correctness of a triple is proportional to $r^\top(e_s \star e_o)$ where $e_s \star e_o$ represents the circular correlation between vectors e_s and e_o and

$$[e_s \star e_o]_k = \sum_{i=0}^{n-1} e_{s_i} e_{o_{(k+i) \bmod n}}$$

- $\star$ is asymmetric, thus HolE allows asymmetric relations
- Logistic loss or pair-wise ranking loss can be used for training
- Algebraically equivalent [Hayashi and Shimbo, 2017] to ComplEx!
- Neither models transitivity, important for instance-of and subtype-of relations in KG

Order embeddings [Vendrov et al., 2015]

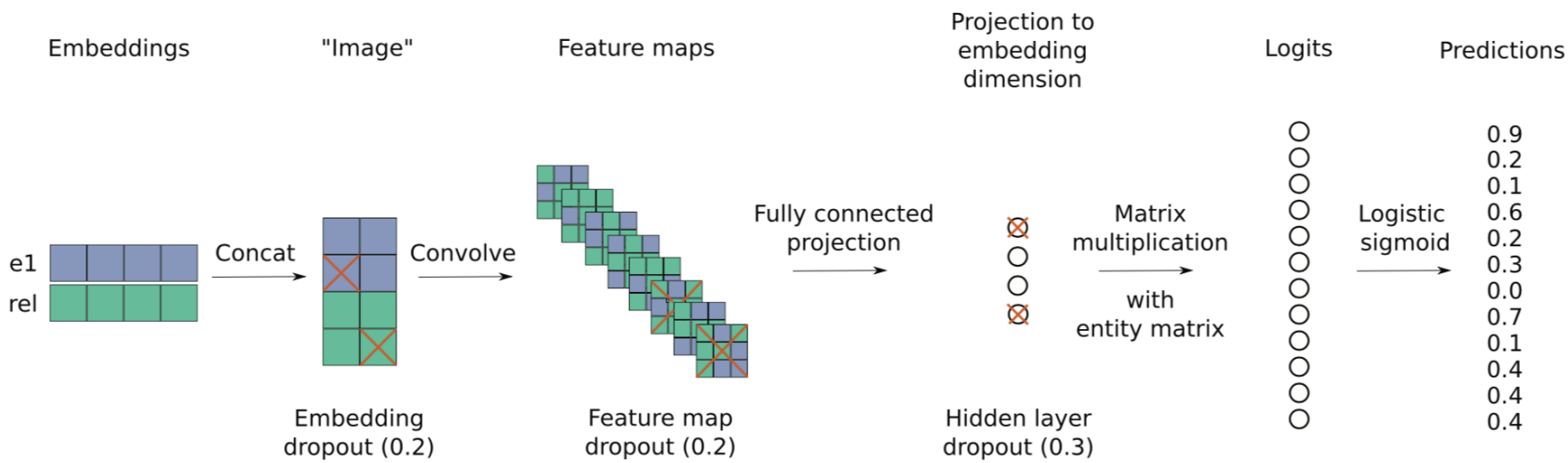
- Specialized to represent partial orders like $e \in t$ and $t_1 \subseteq t_2$, denoted uniformly as $x_1 \prec x_2$
- Embed each x to vector $\boldsymbol{x}$
- If $x_1 \prec x_2$, assert $\boldsymbol{x}_1 \leq \boldsymbol{x}_2$, elementwise
- E.g., by assessing a hinge loss $\|(\boldsymbol{x}_1 - \boldsymbol{x}_2)_+\|_1$, where $(\boldsymbol{a})_+ = (\max\{0, a_i\})$ is an elementwise ReLU
- Negative sampling as usual: pick $x_1 \prec x_2$, perturb either to get $x_1' \not\prec x_2'$; no checking for false negative
- Let $E(x_1, x_2) = \|(\boldsymbol{x}_1 - \boldsymbol{x}_2)_+\|_1$
- Loss function is:

$$\sum_{x_1 \prec x_2} E(x_1, x_2) + \sum_{x_1 \not\prec x_2} \max\{0, \gamma - E(x_1, x_2)\}$$

  where γ is a margin

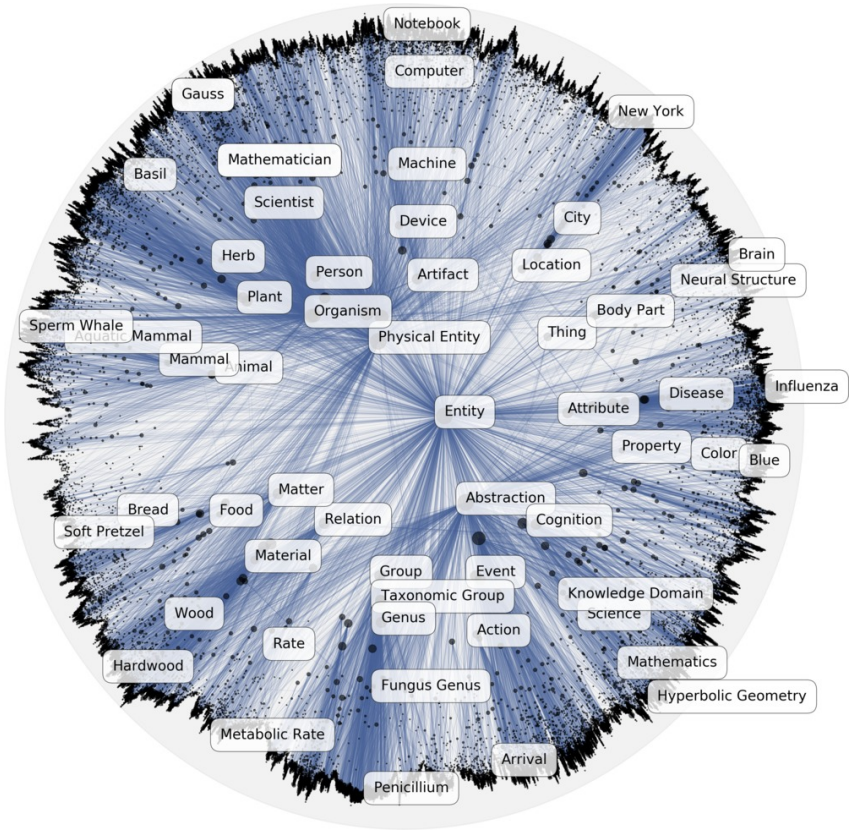
ConvE [Dettmers et al., 2018]

- Uses multi-layer CNN to generate deep and more expressive features as compared to shallow features generated by most of the existing models
- Entities and Relations are modeled as vectors
- Parameter Efficient (uses 8x fewer parameters than DistMult to achieve similar performance)
- Allows cross-dimensional interaction of vectors, similar to HolE



Poincaré Embedding [Nickel and Kiela, 2017]

- Uses Hyperbolic space (n-dimensional Poincaré ball) instead of Euclidean space
- The hyperbolic geometry allows to capture hierarchy and similarity simultaneously
- The structural bias induced by hyperbolic space results in improved generalization performance



The menagerie

Model	Score function $s(e_s, r, e_o)$	
SE	$\|M_{rs}e_s - M_{ro}e_o\|_2$	
TransE	$\|e_s + v_r - e_o\|_2$	
STransE	$\|M_{rs}e_s + v_r - M_{ro}e_o\|_2$	
TransR	$\|M_r e_s + v_r - M_r e_o\|_2$	$M_{rs} = M_{ro} = M_r$
DistMult	(Maximize) $e_s^\top M_r e_o$	M_r diagonal
TransH	$\|e_s - (e_s \cdot p_r)e_s + d_r - e_o + (e_o \cdot p_r)e_o\|_2$	$\|p_r\|_2 = 1$
TransD	$\|(p_r e_s^\top + \mathbb{I})e_s + d_r - (p_r e_o^\top + \mathbb{I})e_o\|_2$	
NTN	$u_r^\top \tanh(e_s^\top \boldsymbol{M}_r e_o + M_{rs}e_s + M_{ro}e_o + b_r)$	$\boldsymbol{M}_r \in \mathbb{R}^{D\times D\times K}$
ConvE	$f(\text{vec}(f([\overline{e_s}; \overline{v_r}] * \omega))W)e_0$	f:activation function

- Who is better than who?
- Just check the publication date!

Performance on generic link prediction

Method	WN18			FB15k		
	MR	**MRR**	**H@10**	**MR**	**MRR**	**H@10**
ConvE	504	**94.2**	**95.5**	**64**	**74.5**	**87.3**
ITransF	**223**		95.2	77		81.4
ComplEx		94.1	94.7		69.2	84
STransE	244		94.7	69		79.7
HolE		93.8	94.9		52.4	73.9
TransR	225		92	77		68.7
TransH	303		86.7	87		64.4
DistMult		82.2	93.6		65.4	82.4
TransE	251	45.4	93.4	125	38.0	47.1

- Yet top predictions show spectacular type errors [Jain et al., 2018][3]

Subject s	**Relation** r	**Gold Object** o	**Prediction**
Howard Leslie Shore	follows-religion	Jewism (religion)	Walk Hard (film)
Spyglass Entertainment	headquarter-located-in	El lay (location)	The Real World (tv)
Les Fradkin	born-in-location	New York (location)	Federico Fellini (person)
Eugene Alden Hackman	studied	Rural Journalism	Loudon Wainwright
Chief Phillips (film)	released-in-region	Yankee land (location)	Akira Isida (person)

[3] Also KG4IR@SIGIR talk

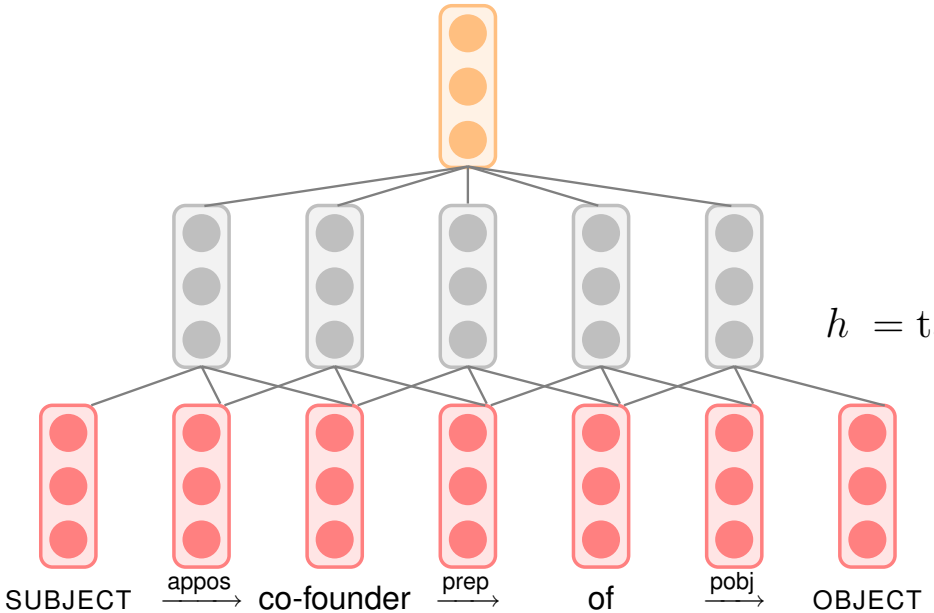
Embeddings from combining KG and text

- TL;DR Use convnet on dependency path to get compositional embedding for r, then combine using DistMult with entity embeddings
- Example dependency paths for (person, founded, organization)

Textual Pattern	Count
SUBJECT $\xrightarrow{\text{appos}}$ founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	12
SUBJECT $\xleftarrow{\text{nsubj}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{appos}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{conj}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	3
SUBJECT $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{nsubj}}$ signed $\xrightarrow{\text{xcomp}}$ establishing $\xrightarrow{\text{dobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xrightarrow{\text{appos}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{nsubj}}$ one $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{nsubj}}$ founded $\xrightarrow{\text{dobj}}$ production $\xrightarrow{\text{conj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{appos}}$ partner $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ founded $\xrightarrow{\text{dobj}}$ production $\xrightarrow{\text{conj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{pobj}}$ by $\xleftarrow{\text{prep}}$ co-founded $\xleftarrow{\text{rcmod}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nn}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	1
SUBJECT $\xrightarrow{\text{dep}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nsubj}}$ helped $\xrightarrow{\text{xcomp}}$ establish $\xrightarrow{\text{dobj}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nsubj}}$ signed $\xrightarrow{\text{xcomp}}$ creating $\xrightarrow{\text{dobj}}$ OBJECT	1

Embeddings from combining KG and text (2)

- Composing text into M of DistMult



$$r = \max\{h\}$$

$$h = \tanh(W^{-1}v_{-1} + W^{0}v + W^{1}v_{+1} + b)$$

$$v = Ve$$

- (Presumably LSTMs have been tried too)
- Model $p(e_o|e_s, r; \Theta) = \dfrac{\exp(f(e_s, r, e_o; \Theta))}{\sum_{e' \in \text{Neg}(e_s, r, ?)} \exp(f(e_s, r, e'; \Theta))}$
- f is the function implemented by the convnets and a final DistMult: $\vec{e}_s^{\top} \operatorname{diag}(r) \vec{e}_o$

Embeddings from combining KG and text (3)

- Two potential issues
  - Although Neg is sampled from all possible negative e', denominator not scaled suitably
  - Positive term not included in denominator
- Θ includes M_r from KG, convnet weights from corpus, and $\vec{e}$
- Overall objective to maximize for each source is

$$L(\mathcal{T};\Theta) = \sum_{(e_s,r,e_o)\in\mathcal{T}} \log p(e_o|e_s,r;\Theta) + \sum_{(e_s,r,e_o)\in\mathcal{T}} \log p(e_s|e_o,r;\Theta)$$

  where $\mathcal{T} \in \{\mathcal{T}_{\text{KG}}, \mathcal{T}_{\text{corpus}}\}$
- Global objective to maximize is

$$L(\mathcal{T}_{\text{KG}};\Theta) + \clubsuit L(\mathcal{T}_{\text{corpus}};\Theta) - \spadesuit \|\Theta\|^2$$

  convnets?

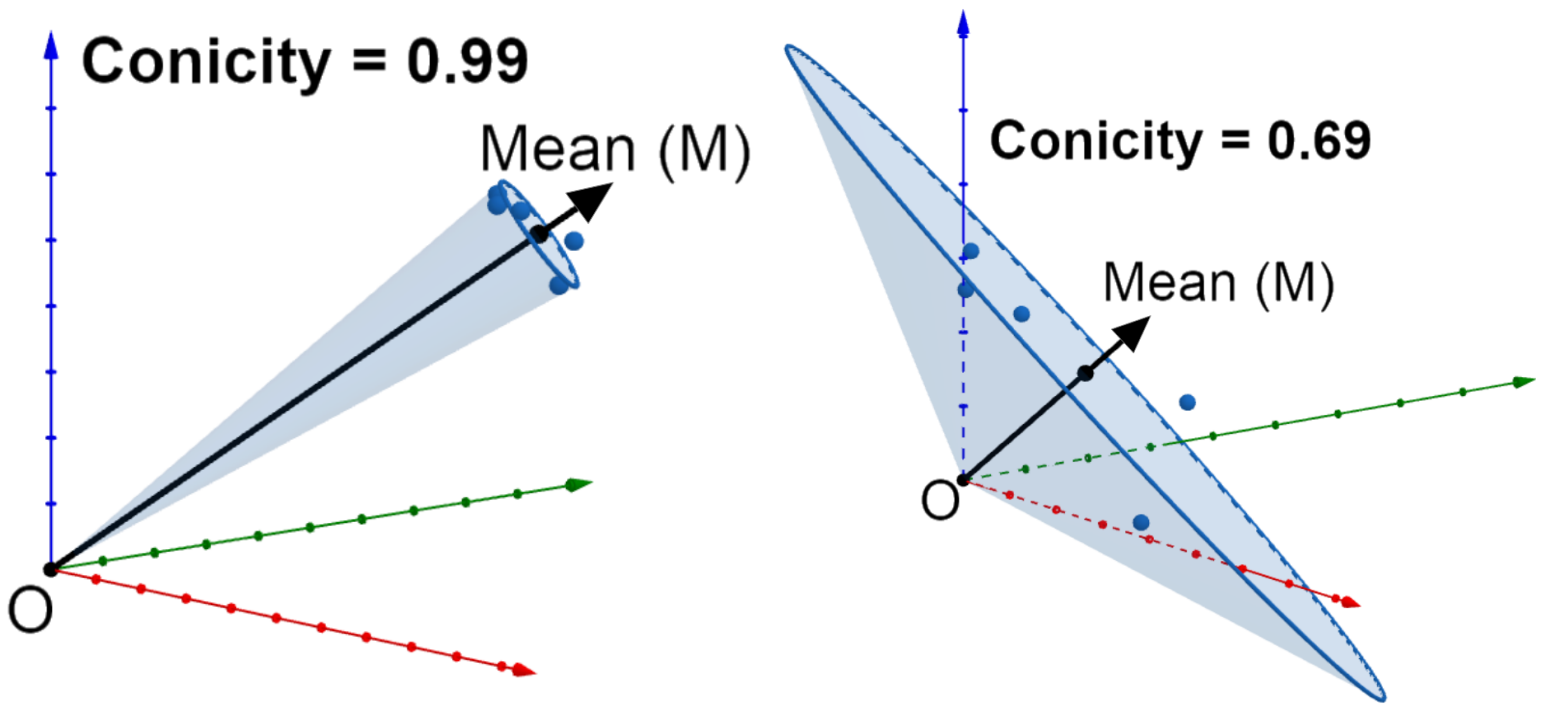
Universal schema [Riedel et al., 2013]

- Another popular technique to combine KG- and corpus-based knowledge inference
- Based on low-rank matrix factorization
- Each row corresponds to an entity pair $\langle e_i, e_j \rangle$
- A column can correspond to
  - A canonical relation r in KG, such as born-in, or
  - A non-canonicalized, textually expressed relation such as "e_i is a native of e_j" or "e_i left his birthplace e_j"
- In the simpler version, let each distinct textual expression be given its own column
  - Separate columns for "is a native of" and "originally hailed from"
- Embedding u_{e_i,e_j} for each entity pair, $v_{\text{born-in}}$ for each canonical relation, and $v_{\text{"is a native of"}}$ for each distinct textually expressed relation
- Matrix factorization expected to make

  $v_{\text{born-in}} \approx v_{\text{"is a native of"}} \approx v_{\text{"originally hailed from"}}$

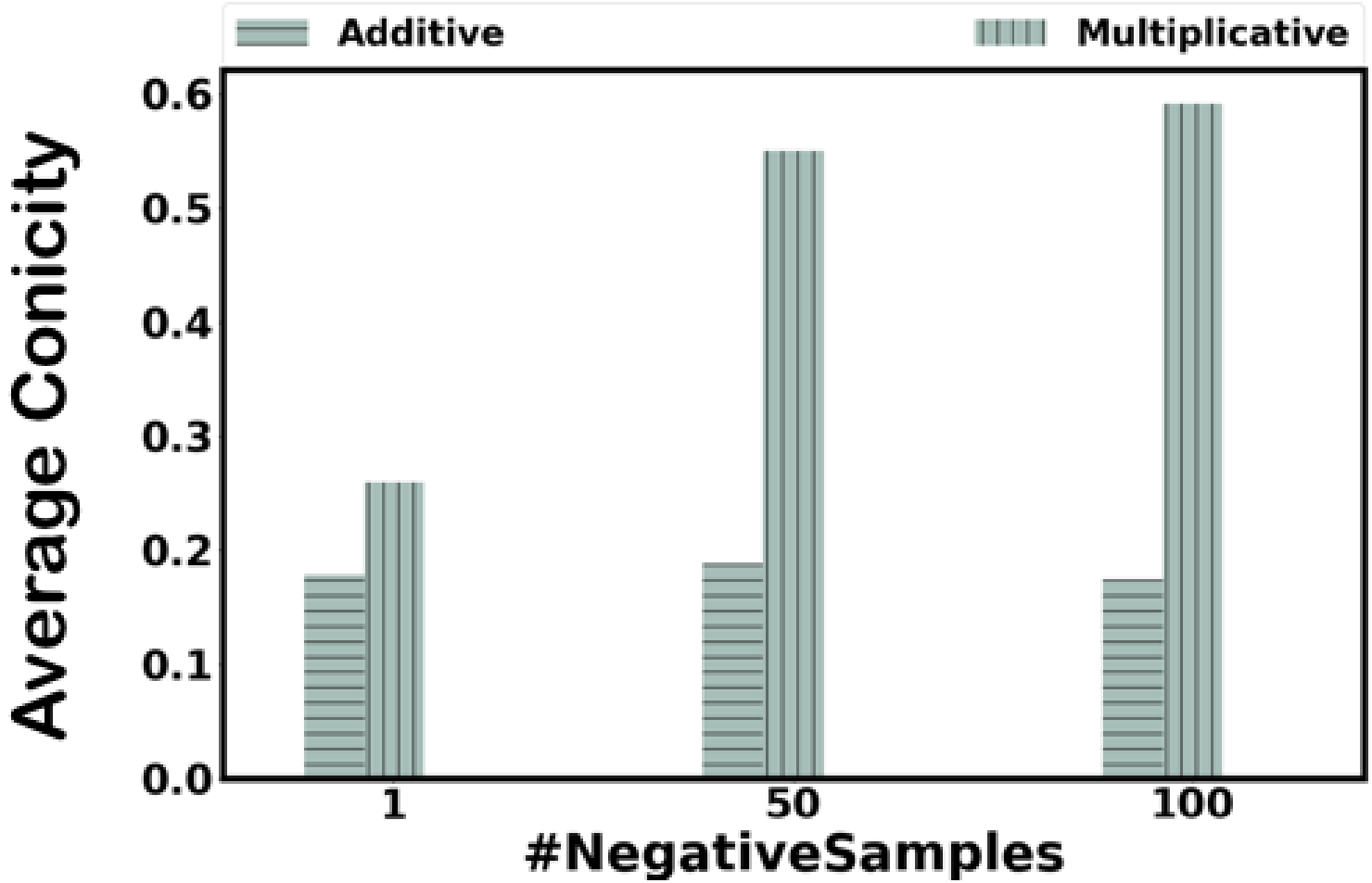
KG Geometry [Chandrahas et al., 2018]

- Studies the geometry (i.e. arrangement of vectors in vector space) of KG embeddings generated by various methods
- Demonstrates a consistent difference between the geometry of Additive Models (e.g. TransE, TransR) and Multiplicative Models (e.g. DistMult, ComplEx) across multiple datasets



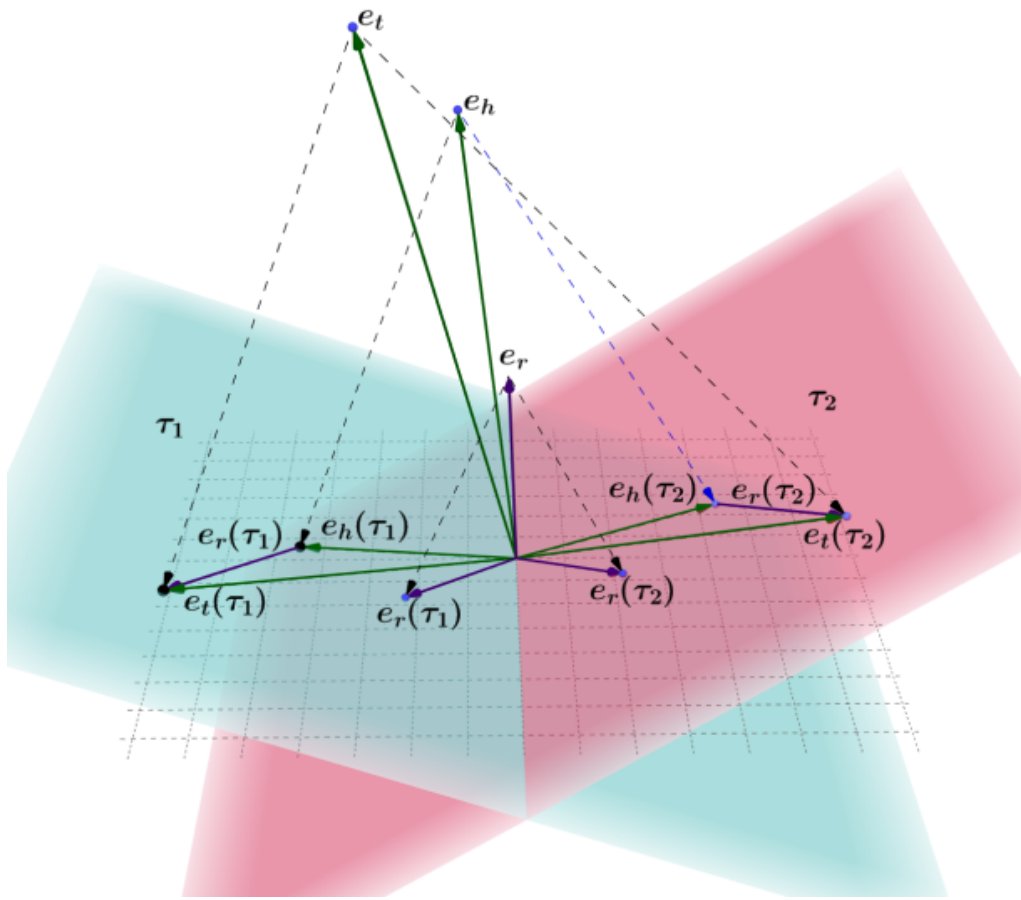
KG Geometry [Chandrahas et al., 2018]

- Geometry of Multiplicative Models are sensitive to #negative-samples while Additive models are not
- Multiplicative models have higher conicity than additive models



HyTE: Temporal KG Embedding [Dasgupta et al., 2018]

- Knowledge (entities and relationships) change over time
- Infuses temporal information into KG Embedding by having separate hyper-planes for different time-intervals
- Use translational models (e.g. TransE) within each hyper-plane



HyTE: Temporal KG Embedding [Dasgupta et al., 2018]

- Allows link as well as temporal scope prediction and outperforms existing methods on YAGO11K and WikiData12k datasets

Dataset	YAGO11K				Wikidata12K			
Metric	Mean Rank		Hits@10(%)		Mean Rank		Hits@10(%)	
	tail	head	tail	head	tail	head	tail	head
Trans-E	504	2020	4.4	1.2	520	740	11.0	6.0
TransH	354	1808	5.8	1.5	423	648	23.7	11.8
HolE	1828	1953	29.4	13.7	734	808	25.0	12.3
t-TransE	292	1692	6.2	1.3	283	413	24.5	14.5
HyTE	**107**	**1069**	**38.4**	**16.0**	**179**	**237**	**41.6**	**25.0**

- The learnt time vectors demonstrate temporal consistency where similar time-intervals form clusters

Outline

13:00-13:15 Overview and motivation

13:15-13:45 Case study: NELL

13:45-14:00 Bootstrapped Entity Extraction

14:00-15:00 Open Relation Extraction & Canonicalization

15:00-15:30 Coffee Break

15:30-16:15 Distantly-supervised Neural Relation Extraction

16:15-16:45 Knowledge Graph Embeddings

16:45-17:00 Conclusion & QA

Timeline: some highlights

$-\infty$ Hidden Markov models

1992 Hearst patterns [Hearst, 1992]

1998 Duality in pattern-relationship extraction [Brin, 1998]

2000 Snowball = DIPRE + confidence scores [Agichtein and Gravano, 2000]

2001 Conditional Random Fields [Lafferty et al., 2001]

2001 Turney's PMI [Turney, 2001]

2001 SemTag and Seeker [Dill et al., 2003]

2000–now Many systems for labeling token spans [Lafferty et al., 2001] or 2d regions [Zhu et al., 2007] with entity types

2001–2002 Search in graph data models [Bhalotia et al., 2002, Agrawal et al., 2002, Hristidis et al., 2003]

2002–2003 Personalized and topic-specific PageRank[Jeh and Widom, 2003, Haveliwala, 2002]

Timeline: some highlights (2)

2004 ObjectRank[Balmin et al., 2004]

2004 KnowItAll $\approx$ Hearst patterns + list extraction + a few more tricks [Etzioni et al., 2004]

2005 Relation extraction via dependency path kernels [Bunescu and Mooney, 2005]

2007 TextRunner [Banko et al., 2007] and ExDB [Cafarella et al., 2007]

2006–2007 Type+proximity search, EntityRank[Chakrabarti et al., 2006, Cheng et al., 2007]

2007–2008 Proximity search in graphs [Chakrabarti, 2007, Sarkar et al., 2008]

2006–2009 Entity disambiguation [Bunescu and Pasca, 2006, Mihalcea and Csomai, 2007, Cucerzan, 2007, Milne and Witten, 2008, Kulkarni et al., 2009]

Timeline: some highlights (3)

2007–2009 Searching the structured-unstructured divide [Kasneci et al., 2008, Suchanek et al., 2007]

2010– Never ending language learning (NELL, @CMU)

2011–2012 MultiR, MIML-RE [Surdeanu et al., 2012a]

2013–2014 Word2vec [Mikolov et al., 2013], GloVE [Pennington et al., 2014]

2014– Continuous knowledge representation for KBC [Bordes et al., 2013]

2011, 2015 Path-ranking algorithm, KG+corpus for relation extraction [Lao and Cohen, 2010, Toutanova et al., 2015]

2004–2006, 2009, 2013–2016 KG+corpus in QA

2016– RL + Relation Extraction

2015– Neural Distant Supervision

2012– Large-scale KG applications

...

Other related recent tutorials

- WSDM 2018 and AAAI 2017 tutorials by Pujara and Singh: https://kgtutorial.github.io/
- SIGIR 2018 tutorial by Chakrabarti: https://goo.gl/vRkwxZ, CIKM 2017 https://goo.gl/A6ZqBq
- Tutorials by Xiang Ren et al.
  - CIKM 2017: https://goo.gl/rpD2Tg
  - NAACL 2018: https://goo.gl/D2HWf1
  - WWW 2018: https://goo.gl/JCwjpw
- WWW 2015 tutorial by Yago group: http://resources.mpi-inf.mpg.de/yago-naga/www2015-tutorial/
- VLDB 2014 tutorial by Suchanek and Weikum: http://resources.mpi-inf.mpg.de/yago-naga/vldb2014-tutorial/

The End

References

E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *ICDL*, pages 85–94, 2000. URL http://www.academia.edu/download/31007490/cucs-033-99.pdf.

S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A system for keyword-based search over relational databases. In *ICDE*, San Jose, CA, 2002. IEEE.

G. Angeli, J. Tibshirani, J. Wu, and C. D. Manning. Combining distant and partial supervision for relation extraction. In *EMNLP Conference*, pages 1556–1567, 2014. URL http://www.anthology.aclweb.org/D/D14/D14-1164.pdf.

A. Balmin, V. Hristidis, and Y. Papakonstantinou. Authority-based keyword queries in databases using ObjectRank. In *VLDB Conference*, Toronto, 2004.

M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In M. M. Veloso, editor, *IJCAI*, pages 2670–2676, 2007. URL http://www.ijcai.org/papers07/Papers/IJCAI07-429.pdf.

G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*. IEEE, 2002.

References (2)

A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *AAAI Conference*, pages 301–306, 2011. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3659/3898.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS Conference*, pages 2787–2795, 2013. URL http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf.

S. Brin. Extracting patterns and relations from the World Wide Web. In P. Atzeni, A. O. Mendelzon, and G. Mecca, editors, *WebDB Workshop*, volume 1590 of *LNCS*, pages 172–183, Valencia, Spain, Mar. 1998. Springer. ISBN 3-540-65890-4. URL http://ilpubs.stanford.edu:8090/421/1/1999-65.pdf.

R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, pages 9–16, 2006. URL http://www.cs.utexas.edu/~ml/papers/encyc-eacl-06.pdf.

References (3)

R. C. Bunescu and R. J. Mooney. A shortest path dependency kernel for relation extraction. In *EMNLP Conference*, pages 724–731. ACL, 2005. doi: http://dx.doi.org/10.3115/1220575.1220666. URL http://acl.ldc.upenn.edu/H/H05/H05-1091.pdf.

M. J. Cafarella, C. Re, D. Suciu, O. Etzioni, and M. Banko. Structured querying of web text: A technical challenge. In *CIDR*, pages 225–234, 2007. URL http://www-db.cs.wisc.edu/cidr/cidr2007/papers/cidr07p25.pdf.

S. Chakrabarti. Dynamic personalized PageRank in entity-relation graphs. In *WWW Conference*, Banff, May 2007. URL http://www.cse.iitb.ac.in/~soumen/doc/netrank/.

S. Chakrabarti, K. Puniyani, and S. Das. Optimizing scoring functions and indexes for proximity search in type-annotated corpora. In *WWW Conference*, pages 717–726, Edinburgh, May 2006. URL http://www.cse.iitb.ac.in/~soumen/doc/www2006i.

References (4)

C. Chandrahas, A. Sharma, and P. Talukdar. Towards understanding the geometry of knowledge graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 122–131. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1012.

T. Cheng, X. Yan, and K. C.-C. Chang. EntityRank: Searching entities directly and holistically. In *VLDB Conference*, pages 387–398, Sept. 2007. URL http://www-forward.cs.uiuc.edu/pubs/2007/entityrank-vldb07-cyc-jul07.pdf.

J. Christensen, Mausam, S. Soderland, and O. Etzioni. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, K-CAP '11, pages 113–120, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0396-5. doi: 10.1145/1999676.1999697. URL http://doi.acm.org/10.1145/1999676.1999697.

S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP Conference*, pages 708–716, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1074.

References (5)

S. S. Dasgupta, S. N. Ray, and P. Talukdar. Hyte: Hyperplane-based temporal embedding for knowledge graph. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

T. Dettmers, M. Pasquale, S. Pontus, and S. Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, pages 1811–1818, February 2018. URL https://arxiv.org/abs/1707.01476.

S. Dill et al. SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation. In *WWW Conference*, pages 178–186, 2003.

O. Etzioni, M. Cafarella, et al. Web-scale information extraction in KnowItAll. In *WWW Conference*, New York, 2004. ACM. URL http://www.cs.washington.edu/research/knowitall/papers/www-paper.pdf.

J. Feng. Reinforcement Learning for Relation Extraction from Noisy Data. *AAAI*, pages 1–10, 2018. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/AAAI2018Denoising.pdf.

References (6)

L. Galárraga, G. Heitz, K. Murphy, and F. M. Suchanek. Canonicalizing open knowledge bases. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1679–1688, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. doi: 10.1145/2661829.2662073. URL http://doi.acm.org/10.1145/2661829.2662073.

L. A. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek. Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 413–422, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488425. URL http://doi.acm.org/10.1145/2488388.2488425.

X. Han. Global Distant Supervision for Relation Extraction. *AAAI*, pages 2950–2956, 2016.

T. H. Haveliwala. Topic-sensitive PageRank. In *WWW Conference*, pages 517–526, 2002. URL http://www2002.org/CDROM/refereed/127/index.html.

References (7)

K. Hayashi and M. Shimbo. On the equivalence of holographic and complex embeddings for link prediction. *arXiv preprint arXiv:1702.05563*, 2017. URL https://arxiv.org/pdf/1702.05563.

M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics*, volume 14, pages 539–545, 1992. URL http://www.aclweb.org/website/old_anthology/C/C92/C92-2082.pdf.

R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL Conference*, pages 541–550, 2011. URL http://anthology.aclweb.org/P/P11/P11-1055.pdf.

V. Hristidis, L. Gravano, and Y. Papakonstantinou. Efficient IR-style keyword search over relational databases. In *VLDB Conference*, pages 850–861, 2003. URL http://www.db.ucsd.edu/publications/VLDB2003cr.pdf.

P. Jain, P. Kumar, Mausam, and S. Chakrabarti. Type-sensitive knowledge base inference without explicit type supervision. In *ACL Conference*, 2018.

S. Jat, S. Khandelwal, and P. Talukdar. Improving distantly supervised relation extraction using word and entity based attention. *CoRR*, abs/1804.06987, 2018. URL http://arxiv.org/abs/1804.06987.

References (8)

G. Jeh and J. Widom. Scaling personalized web search. In *WWW Conference*, pages 271–279, 2003. URL https://goo.gl/oAZZLq.

G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL Conference*, pages 687–696, 2015. URL http://www.aclweb.org/anthology/P/P15/P15-1067.pdf.

G. Kasneci, F. M. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath, and G. Weikum. NAGA: harvesting, searching and ranking knowledge. In *SIGMOD Conference*, pages 1285–1288. ACM, 2008. ISBN 978-1-60558-102-6. doi: http://doi.acm.org/10.1145/1376616.1376756. URL http://www.mpi-inf.mpg.de/~kasneci/naga/.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of Wikipedia entities in Web text. In *SIGKDD Conference*, pages 457–466, 2009. URL http://www.cse.iitb.ac.in/~soumen/doc/CSAW/.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001. URL ftp://ftp.cis.upenn.edu/pub/datamining/public_html/ReadingGroup/papers/crf.pdf.

References (9)

N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(1):53–67, Oct. 2010. ISSN 0885-6125. doi: 10.1007/s10994-010-5205-8. URL http://dx.doi.org/10.1007/s10994-010-5205-8.

Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. Neural relation extraction with selective attention over instances. In *ACL Conference*, pages 2124–2133, August 2016a. URL http://www.aclweb.org/anthology/P16-1200.

Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany, August 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1200.

M. Mausam. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 4074–4077. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3061053.3061220.

References (10)

R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242, 2007. URL http://portal.acm.org/citation.cfm?id=1321440.1321475.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS Conference*, pages 3111–3119, 2013. URL https://goo.gl/x3DTzS.

D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM*, pages 509–518, 2008. URL http://www.cs.waikato.ac.nz/~dnk2/publications/CIKM08-LearningToLinkWithWikipedia.pdf.

B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *NAACL Conference*, pages 777–782, 2013. URL http://www.anthology.aclweb.org/N/N13/N13-1095.pdf.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL Conference*, pages 1003–1011, 2009. URL http://www.aclweb.org/anthology/P09-1113.

References (11)

T. Mitchell, W. Cohen, E. Hruscha, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohammad, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *AAAI*, 2015. URL http://www.cs.cmu.edu/~wcohen/pubs.html. : Never-Ending Learning in AAAI-2015.

D. Movshovitz-Attias and W. W. Cohen. Kb-lda: Jointly learning a knowledge base of hierarchy, relations, and facts. In *ACL*, 2015.

M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1): 11–33, Jan 2016a. ISSN 0018-9219. doi: 10.1109/JPROC.2015.2483592.

M. Nickel, L. Rosasco, T. A. Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI Conference*, pages 1955–1961, 2016b. URL https://arxiv.org/abs/1510.04935.

References (12)

M. Nimishakavi, U. S. Saini, and P. Talukdar. Relation schema induction using tensor factorization with side information. In *Empirical Methods in Natural Language Processing*, 2016.

M. Nimishakavi, M. Gupta, and P. Talukdar. Higher-order relation schema induction using tensor factorization with back-off and aggregation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1575–1584. Association for Computational Linguistics, 2018.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *EMNLP Conference*, volume 14, pages 1532–1543, 2014. URL http://www.emnlp2014.org/papers/pdf/EMNLP2014162.pdf.

P. Qin, W. Xu, and W. Y. Wang. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. *Transactions of the Association for Computational Linguistics*, 5, 2018. URL http://arxiv.org/abs/1805.09929.

References (13)

C. Quirk and H. Poon. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1171–1182, 2017. URL https://aclanthology.info/papers/E17-1110/e17-1110.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62 (1-2):107–136, Feb. 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-5833-1. URL http://dx.doi.org/10.1007/s10994-006-5833-1.

S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. Relation extraction with matrix factorization and universal schemas. In *NAACL Conference*, pages 74–84, 2013. URL http://www.anthology.aclweb.org/N/N13/N13-1008.pdf.

References (14)

S. Saha, H. Pal, and Mausam. Bootstrapping for numerical open ie. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-2050. URL http://www.aclweb.org/anthology/P17-2050.

S. Sarawagi. Information extraction. *FnT Databases*, 1(3), 2008. URL http://www.cse.iitb.ac.in/~sunita/papers/ieSurvey.pdf.

P. Sarkar, A. W. Moore, and A. Prakash. Fast incremental proximity search in large graphs. In *ICML*, pages 896–903, 2008. URL http://icml2008.cs.helsinki.fi/papers/565.pdf.

F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *WWW Conference*, pages 697–706. ACM Press, 2007. URL http://www2007.org/papers/paper391.pdf.

M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. Multi-instance multi-label learning for relation extraction. In *EMNLP Conference*, pages 455–465, 2012a. URL http://anthology.aclweb.org/D/D12/D12-1042.pdf.

References (15)

M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics, 2012b.

P. P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1473–1481, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858830.

K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon. Representing text for joint embedding of text and knowledge bases. In *EMNLP Conference*, pages 1499–1509, 2015. URL https://www.aclweb.org/anthology/D/D15/D15-1174.pdf.

T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080, 2016. URL http://arxiv.org/abs/1606.06357.

References (16)

P. D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML*, 2001.

S. Vashishth, P. Jain, and P. Talukdar. Cesi: Canonicalizing open knowledge bases using embeddings and side information. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1317–1327, Republic and Canton of Geneva, Switzerland, 2018a. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186030. URL https://doi.org/10.1145/3178876.3186030.

S. Vashishth, R. Joshi, S. S. Prayaga, C. Bhattacharyya, and P. Talukdar. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '18)*, Brussels, Belgium, Oct 31-Nov 4 2018b.

I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. URL https://arxiv.org/pdf/1511.06361.

References (17)

Q. Xie, X. Ma, Z. Dai, and E. Hovy. An interpretable knowledge transfer model for knowledge base completion. *arXiv preprint arXiv:1704.05908*, 2017. URL https://arxiv.org/pdf/1704.05908.pdf.

Y. Yaghoobzadeh, H. Adel, and H. Schütze. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/E17-1111.

D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP Conference*, pages 1753–1762, 2015. URL http://www.aclweb.org/anthology/D15-1203.

J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen. Dynamic hierarchical Markov random fields and their application to Web data extraction. In *ICML*, pages 1175–1182, 2007. URL http://www.machinelearning.org/proceedings/icml2007/papers/215.pdf.