Organizing Web Information

CS 728

Soumen Chakrabarti
IIT Bombay
http://www.cse.iitb.ac.in/~soumen/

From coarse NER to fine types

FIGER type catalog (112 fine types)

person	doctor
actor	engineer
architect	monarch
artist	musician
athlete	politician
author	religious_leader
coach	soldier
director	terrorist

organization	terrorist_organization
airline	government_agency
company	government
educational_institution	political_party
fraternity_sorority	educational_department
sports_league	military
sports_team	news_agency

location	body_of_water
city	island
country	mountain
county	glacier
province	astral_body
railway	cemetery
road	park
bridge	

product	camera
engine	mobile_phone
airplane	computer
car	software
ship	game
spacecraft	instrument
train	weapon

art	written_work
film	newspaper
play	music

event	military_conflict
attack	natural_disaster
election	sports_event
protest	terrorist_attack

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

Fine type tagging: Motivation

- Suppose John Smith is a cricket player not yet in Wikipedia
- But mentioned in local news about county cricket
- Query is "Who took four wickets in one over last year against Birmingham?"
- Potential evidence passage[3] is "Birmingham crashed out of the match after losing four wickets to Smith in a single over last month."
- Goal is to collect John Smith as a (strong) candidate, for which we must know that Smith refers to a cricketer[4]
- Experience suggests (thousands of) finer types better for QA than (hundreds of) fine types, but hard to infer from context

[3] Would be very nice to also collect evidence of four wickets from "Alan and Boyd were bowled out by the first two balls from Smith; Ray and Tony were caught out before the over was done."

[4] Must also know that who is asking for a cricketer, not, e.g., a politician, a process called answer/target type inference.

Type tagging: basic idea

- Efficiently produce training data: text with entity mention spans marked out, with type(s) of entities provided as labels
  - Nobody scored as many goals in one match as Messi in 2004.
  - Type of Messi is `/person/athlete`
- Source: Wikipedia links to other Wikipedia pages corresponding to entities
- Collect features from mention context
  - scored, goals, match
- Find types to which these entities belong — these are labels
- (Caveat: Not all these types may be active in a mention context)
- Train a multi-class, multi-label classifier
- At test time, use a B-I-O CRF to locate mention segments
- For each mention, collect features from context
- Predict one or more types using multi-class, multi-label classifier

FIGER system and features

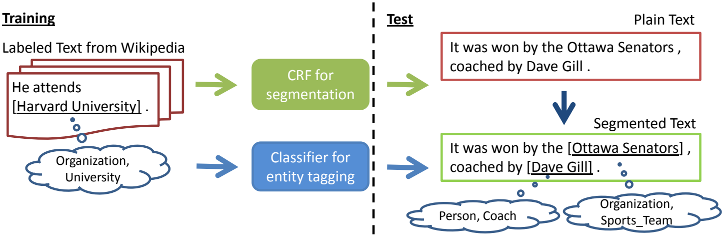


Figure 1: System architecture of FIGER.

Feature	Decription	Example
Tokens	The tokens of the segment.	"Eton"
Word Shape	The word shape of the tokens in the segment.	"Aa" for "Eton" and "A0" for "CS446".
Part-of-Speech tags	The part-of-speech tags of the segment.	"NNP"
Length	The length of the segment.	1
Contextual unigrams	The tokens in a contextual window of the segment.	"victory", "for", "."
Contextual bigrams	The contextual bigrams including the segment.	"victory for", "for Eton" and "Eton ."
Brown clusters	The cluster id of each token in the segment (using the first 4, 8 and 12-bit prefixes).	"4_1110", "8_11100111", etc.
Head of the segment	The head of the segment following the rules by Collins (1999).	"HEAD_Eton"
Dependency	The Stanford syntactic dependency (De Marneffe, MacCartney, and Manning 2006) involving the head of the segment.	"prep_for:seal:dep"
ReVerb patterns	The frequent lexical patterns as meaningful predicates	"seal_victory_for:dep"

Google fine types and baseline system

PERSON

- **artist**
  - actor
  - author
  - director
  - music
- **education**
  - student
  - teacher
- **athlete**
- **business**
- **coach**
- **doctor**
- **legal**
- **military**
- **political figure**
- **religious leader**
- **title**

LOCATION

- **structure**
  - airport
  - government
  - hospital
  - hotel
  - restaurant
  - sports facility
  - theatre
- **geography**
  - body of water
  - island
  - mountain
- **transit**
  - bridge
  - railway
  - road
- **celestial**
- **city**
- **country**
- **park**

ORGANIZATION

- **company**
  - broadcast
  - news
- **education**
- **government**
- **military**
- **music**
- **political party**
- **sports league**
- **sports team**
- **stock exchange**
- **transit**

OTHER

- **art**
  - broadcast
  - film
  - music
  - stage
  - writing
- **event**
  - accident
  - election
  - holiday
  - natural disaster
  - protest
  - sports event
  - violent conflict
- **health**
  - malady
  - treatment
- **award**
- **body part**
- **currency**
- **language**
  - programming language
- **living thing**
  - animal
- **product**
  - camera
  - car
  - computer
  - mobile phone
  - software
  - weapon
- **food**
- **heritage**
- **internet**
- **legal**
- **religion**
- **scientific**
- **sports & leisure**
- **supernatural**

- Minor tweaks to FIGER types
- Improvements in collecting labeled data

Google fine types and baseline system (2)

- Enhanced classification
- Training data expected to have extraneous labels
  - Entity Obama is-a politician, (ex-) POTUS, lawyer, book author, parent, ...
  - In a given context, one or few types may be 'active'
  - But training instance produced with all type labels
- To mitigate problems from extraneous labels, use weighted approximate rank pairwise (WARP) loss
- Features[5] (e.g. for "... who Barack H. Obama first picked ...")

Feature	Description	Example
Head	The syntactic head of the mention phrase	"Obama"
Non-head	Each non-head word in the mention phrase	"Barack", "H."
Cluster	Word cluster id for the head word	"59"
Characters	Each character trigram in the mention head	":ob", "oba", "bam", "ama", "ma:"
Shape	The word shape of the words in the mention phrase	"Aa A. Aa"
Role	Dependency label on the mention head	"subj"
Context	Words before and after the mention phrase	"B:who", "A:first"
Parent	The head's lexical parent in the dependency tree	"picked"
Topic	The most likely topic label for the document	"politics"

[5] "Washington sat on his favorite Barcelona and opened a Newcastle."

Embedding type labels with WARP loss

- Mention contexts represented as x
- A common situation is $x \in \mathbb{R}^D$, for which we choose embedding $f(x) = \boldsymbol{A}x \in \mathbb{R}^H$, where $\boldsymbol{A} \in \mathbb{R}^{H \times D}$
- Want to exploit related types by embedding each type to a vector; similar types expected to embed to similar vectors
- Let $\boldsymbol{\delta}_t$ is the 1-hot vector for t
- Let the tth column of matrix $\boldsymbol{B} \in \mathbb{R}^{H \times T}$ represent the H-dimensional embedding of type t
- I.e., we can use notation $g(t) = \boldsymbol{B}\boldsymbol{\delta}_t$ as the embedding $g(t) \in \mathbb{R}^H$
- The score of a single type label t for context x is $s_t(x) = f(x) \cdot g(\boldsymbol{\delta}_t)$
- Multiple type labels may be valid in both train and test instances

Embedding type labels with WARP loss (2)

- ith labeled instance is $(x_i, \boldsymbol{y}_i)$ where $\boldsymbol{y}_i$ represents a label set, possibly as a few-hot vector in $\{0,1\}^T$
- Exact inference must explore all 2^T label subsets: $\hat{\boldsymbol{y}} = \operatorname{argmax}_{\boldsymbol{y}} f(x) \cdot g(\boldsymbol{y})$
- To avoid high inference cost, cast as label ranking
- Overall score vector $\boldsymbol{s}(x) = (\ldots, s_t(x), \ldots) \in \mathbb{R}^T$
- Goal is to rank all correct labels before any incorrect one
- Loss on instance $x_i, \boldsymbol{y}_i$ is some function of the rank(s) of the correct label(s) in list of types sorted by decreasing score
- Let $\operatorname{rank}(t, \boldsymbol{s}(x))$ be the rank of label t in sorted list

$$\operatorname{rank}(t, \boldsymbol{s}(x)) = \sum_{y' \neq y} \mathbb{I}(s_{y'}(x) \geq s_y(x))$$

- For a single correct t, we can minimize the above rank

Embedding type labels with WARP loss (3)

- For multiple correct ts, there are various options to combine their ranks, e.g., sum
- For instance $x_i, \boldsymbol{y}_i$, consider good type $t \in \boldsymbol{y}_i$, bad type $t' \notin \boldsymbol{y}_i$
- RankSVM loss for such a pair would be $\max\{0, 1 + s_{t'}(x) - s_t(x)\}$
- To incorporate the rank signal of t, define overall WARP loss

$$\sum_{t \in \boldsymbol{y}_i} \sum_{\bar{t} \notin \boldsymbol{y}_i} \mathcal{R}(\operatorname{rank}(t, \boldsymbol{s}(x))) \max\{0, 1 + s_{t'}(x) - s_t(x)\}$$

- Here $\mathcal{R}$ transforms rank into weight; for precision at k, we can use $\mathcal{R} = \sum_{1 \le i \le k} 1/i$
- Not convex

Kernel WSABIE

- Earlier, $s_t(x) = (Ax) \cdot (B\boldsymbol{\delta}_t) = x^\top (A^\top B)\,\boldsymbol{\delta}_t$
- Where $Ax \in \mathbb{R}^H$ and $B\boldsymbol{\delta}_t \in \mathbb{R}^H$
- A and B appear in only the form $A^\top B \in \mathbb{R}^{D\times T}$, but it is constrained to have rank at most H as a form of regularization
- Despite this, observed noisy "fill" in this matrix while training
- Let $P \circ Q$ be the elementwise product of two matrices, i.e., $(P \circ Q)[d,t] = P[d,t]\,Q[d,t]$
- Google system uses $K \in \{0,1\}^{D\times T}$ as a feature selection or additional noise reduction mechanism

$$s_t(x) = x^\top \big(K \circ (A^\top B)\big)\,\boldsymbol{\delta}_t$$

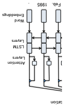
- If $A[:,d]$ is among the 200 nearest neighbors of $B[:,t]$, set $K[d,t] = 1$, and 0 otherwise
- K updated after every iteration (mini-batch?)

Google fine-type system #2 performance

Method	P	R	F1
Ling and Weld (2012)	–	–	69.30
WSABIE	81.85	63.75	71.68
K-WSABIE	**82.23**	**64.55**	**72.35**

Table 4: Precision (P), Recall (R), and F1-score on the FIGER dataset for three competing models. We took the F1 score from Ling and Weld's best result (no precision and recall numbers were reported). The improvements for WSABIE and K-WSABIE over the baseline are statistically significant ($p < 0.01$).

Bi-LSTM fine-type tagger



She got a Ph.D from New York in Feb. 1995.

- Bi-LSTM on left and right context
- Average of word vectors of mention
- +Attention

Bi-LSTM fine-type tagger details

- Let mention words be $M = \{m\}$ with corresponding pretrained (focus) word vectors $u(m)$ from word2vec or GloVe
- Mention vector is designed as $v_m = (1/|M|) \sum_{m \in M} u(m)$
- Suppose we take C words of context from left and right
- Rightmost state from left context LSTM is $\overrightarrow{h}^{\ell}_{C}$
- Leftmost state from right context LSTM is $\overleftarrow{h}^{r}_{1}$
- Context vector is designed as $v_c = \begin{bmatrix} \overrightarrow{h}^{\ell}_{C} \\ \overleftarrow{h}^{r}_{1} \end{bmatrix}$
- Each type t is predicted with

$$\Pr(t|\text{mention, context}) = \sigma\left(W_t \begin{bmatrix} v_m \\ v_c \end{bmatrix}\right)$$

Computing v_c with attention

Sentence	Prediction
... ... The film is a remake of [Secrets (1924)] , a silent film starring Norma Talmadge	/film 0.986 /art 0.982
The film is a remake of Secrets (1924) , a silent film starring [Norma Talmadge]	/person 0.999 /actor 0.987
... ... The festival brought together the foremost filmmakers , including Francois Truffaut , [Roman Polanski] , Robert Enrico , and others	/person 1.00 /director 0.963 /author 0.958 /artist 0.950 /actor 0.871
... ... Jim Hodges , the Democratic nominee , handily defeated Republican Governor [David Beasley] to become the 114th governor of South Carolina	/person 1.00 /politician 0.983
She is best known for roles in various TV Dramas and tokusatsu shows such as [Ultraseven X] and Kamen Rider Kiva	/broadcats_program 0.892

$$e_i^\ell = \tanh\left(W_e \begin{bmatrix} \overrightarrow{h}_i^\ell \\ \overleftarrow{h}_i^\ell \end{bmatrix}\right)$$

L-R & R-L states from left context

$$e_i^r = \cdots$$

L-R & R-L states from right context

$$\tilde{a}_i^\ell = \exp\left(W_a e_i^\ell\right)$$

Attend to important context words

$$\tilde{a}_i^r = \cdots$$

$$a_i^\ell = \frac{\tilde{a}_i^\ell}{\sum_{i=1}^C (\tilde{a}_i^\ell + \tilde{a}_i^r)}$$

Normalize attention over left context

$$v_c = \sum_{i=1}^C a_i^\ell \begin{bmatrix} \overrightarrow{h}_i^\ell \\ \overleftarrow{h}_i^\ell \end{bmatrix} + a_i^r \begin{bmatrix} \overrightarrow{h}_i^r \\ \overleftarrow{h}_i^r \end{bmatrix}$$

Redefined context representation

LSTM and attention results

Models	P	R	F1
Ling and Weld (2012)	-	-	69.30
Yogatama et al. (2015)	**82.23**	64.55	72.35
Averaging Encoder	68.63	69.07	68.65
LSTM Encoder	72.32	70.36	71.34
Attentive Encoder	73.63	**76.29**	**74.94**

Table 1: Loose Micro Precision (P), Recall (R), and F1-score on the test set

Models	Strict	Loose Macro	Loose Micro
Ling and Weld (2012)	52.30	69.90	69.30
Yogatama et al. (2015)	-	-	72.25
Averaging Encoder	51.89	72.24	68.65
LSTM Encoder	55.60	73.95	71.34
Attentive Encoder	**58.97**	**77.96**	**74.94**

Table 2: Strict, Loose Macro and Loose Micro F1-scores

Reducing (type) label noise [9]

- Fine type training data in the form of spans directly gold-labeled with types is rare
- Wikipedia has millions of pages of text with gold mentions of entities
- Wikipedia, DBpedia, Freebase, WikiData, ... have type hierarchies from which we can get all types that contain an entity
- However, most of these types are not relevant at any given mention of the entity
- Training all these types using this textual context would pollute the type models
- Notation: entity e, with mention contexts $C_e = \{c_{ei}\}$ (if e is understood, will drop it)
- e is a member of types in T_e, specified by KG
- I.e., each e associated with $\boldsymbol{y}_e$, a few-hot vector of types

Reducing (type) label noise [9] (2)

- Less realistic to assume per-context gold labels (except to eval fine-type system)
- Each mention context is an instance
- I.e., each entity is associated with multiple instances
- In general each entity has multiple valid labels (types)
- Therefore, a multi-instance multi-label (MIML) setting
- Each context associate with ______ (one/more) types?

MIML approach to fine typing

- Each context c_i will be represented by a fixed-size vector $\boldsymbol{c}_i \in \mathbb{R}^H$ (defined later)
- A first-cut per-mention predictor is a logistic regression: $\Pr(t|c_i) = \sigma(\boldsymbol{w}_t \cdot \boldsymbol{c}_i + b_t)$
- Note multiple t can have score close to 1
- Next, we aggregate in various ways over contexts
- MIML-MAX: Each type $t \in T_e$ is supported by one best context: $\Pr(t|e) = \max_{c \in C_e} \Pr(t|c)$
- Ignores all smaller endorsements
- MIML-AVG: $\Pr(t|e) = \frac{1}{|C_e|} \sum_{c \in C_e} \Pr(t|c)$
- Binary cross entropy BCE$(y, y') = -y \log y' - (1-y) \log(1-y')$
- All $\boldsymbol{w}_t$s can be trained using cross-entropy loss $L(\{\boldsymbol{w}_t\}) = \sum_e \sum_t \text{BCE}\big(y_{et}, \Pr(t|e; \boldsymbol{w}_t)\big)$

MIML approach to fine typing (2)

- MIML-ATT: Aggregate with attention over contexts
- Apart from $\boldsymbol{w}_t$, associate each t with another vector $\boldsymbol{v}_t$
- Mention contexts of entity e compete for attention:
  $\alpha_{i,t} = \dfrac{\exp(\boldsymbol{c}_i \cdot \boldsymbol{v}_t)}{\sum_{i'} \exp(\boldsymbol{c}_{i'} \cdot \boldsymbol{v}_t)}$
- Now we build an attention-weighted context representation:
  $\boldsymbol{a}_t = \sum_i \alpha_{i,t} \boldsymbol{c}_i$
- Use $\boldsymbol{a}_t$ in place of $\boldsymbol{c}_i$ before: $\Pr(t|e) = \sigma(\boldsymbol{w}_t \cdot \boldsymbol{a}_t + b_t)$
- Loss as before
- Additional "deepness": $\alpha_{i,t} = \dfrac{\exp(\boldsymbol{c}_i^\top \boldsymbol{M} \boldsymbol{v}_t)}{\sum_{i'} \exp(\boldsymbol{c}_{i'}^\top \boldsymbol{M} \boldsymbol{v}_t)}$, where $\boldsymbol{M}$ measures the similarity between context and $\boldsymbol{v}_t$

Context representation $\boldsymbol{c}_i$ using convnet

- At the input, read word embeddings
- Apply narrow convnets separately to left and right context of mention to get $\phi_\ell(c), \phi_r(c)$
- Concatenate into $\phi(c)$ and compute $\boldsymbol{c} = \tanh(\boldsymbol{S}\phi(c))$ where $\boldsymbol{S}$ is more model weights
- So overall we have these model weights:
  - Global $\boldsymbol{M}, \boldsymbol{S}$
  - Global weights in convnet ϕ
  - $\boldsymbol{w}_t, \boldsymbol{v}_t, b_t$ for each type
  - Word embeddings (if fine tuned after pretraining)
- Between $\boldsymbol{w}_t, \boldsymbol{v}_t$, is there a usable/interpretable representation of type t?
- (How) do they relate to entity embeddings as in ent2vec?

Noise mitigation results

	$P@1$ all	F_1 all	F_1 head	F_1 tail	MAP
1 MLP	74.3	69.1	74.8	52.5	42.1
2 MLP+MIML-MAX	74.7	59.2	50.7	46.8	41.3
3 MLP+MIML-AVG	77.2	70.6	74.9	56.2	45.0
4 MLP+MIML-MAX-AVG	75.2	71.2	76.4	56.0	47.1
5 MLP+MIML-ATT	81.0	72.0	76.9	59.1	48.8
6 CNN	78.4	72.2	77.3	56.3	47.6
7 CNN+MIML-MAX	78.6	62.2	53.5	49.7	46.6
8 CNN+MIML-AVG	80.8	73.5	77.7	59.2	50.4
9 CNN+MIML-MAX-AVG	79.9	74.3	79.2	59.8	53.3
10 CNN+MIML-ATT	83.4	75.1	79.4	62.2	55.2
11 EntEmb	80.8	73.3	79.9	57.4	56.6
12 FIGMENT	81.6	74.3	80.3	60.1	57.0
13 CNN+MIML-ATT+EntEmb	**85.4**	**78.2**	**83.3**	**66.2**	**64.8**

- ClueWeb with FACC1 entity annotations
- Freebase entities mapped to 102 FIGER types
- 4.3 million contexts
- Head means > 100, tail < 5 mentions

References

[1] D. Freitag and A. McCallum, "Information extraction using HMMs and shrinkage," in *Papers from the AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999, pp. 31–36.

[2] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, no. Sep., pp. 1453–1484, 2005. [Online]. Available: http://ttic.uchicago.edu/~altun/pubs/TsoJoaHofAlt-JMLR.pdf

[3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[4] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *HLT-NAACL*, 2003, pp. 134–141. [Online]. Available: http://acl.ldc.upenn.edu/N/N03/N03-1028.pdf

[5] X. Ling and D. S. Weld, "Fine-grained entity recognition." in *AAAI*, 2012. [Online]. Available: http://xiaoling.github.io/pubs/ling-aaai12.pdf

References (2)

[6] D. Gillick, N. Lazic, K. Ganchev, J. Kirchner, and D. Huynh, "Context-dependent fine-grained entity type tagging," *arXiv preprint arXiv:1412.1820*, 2014. [Online]. Available: https://arxiv.org/pdf/1412.1820.pdf

[7] D. Yogatama, D. Gillick, and N. Lazic, "Embedding methods for fine grained entity type classification," in *ACL Conference*, 2015, pp. 26–31. [Online]. Available: http://anthology.aclweb.org/P/P15/P15-2048.pdf

[8] S. Shimaoka, P. Stenetorp, K. Inui, and S. Riedel, "An attentive neural architecture for fine-grained entity type classification," *arXiv preprint arXiv:1604.05525*, 2016. [Online]. Available: https://arxiv.org/pdf/1604.05525.pdf

[9] Y. Yaghoobzadeh, H. Adel, and H. Schütze, "Noise mitigation for neural entity typing and relation extraction," *arXiv preprint arXiv:1612.07495*, 2016. [Online]. Available: https://arxiv.org/pdf/1612.07495.pdf

References (3)

[10] S. Dill *et al.*, "SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation," in *WWW Conference*, 2003, pp. 178–186.

[11] R. Mihalcea and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," in *CIKM*, 2007, pp. 233–242. [Online]. Available: http://portal.acm.org/citation.cfm?id=1321440.1321475

[12] R. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation," in *EACL*, 2006, pp. 9–16. [Online]. Available: http://www.cs.utexas.edu/~ml/papers/encyc-eacl-06.pdf

[13] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," in *EMNLP Conference*, 2007, pp. 708–716. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1074

[14] J. Hoffart *et al.*, "Robust disambiguation of named entities in text," in *EMNLP Conference*. Edinburgh, Scotland, UK: SIGDAT, Jul. 2011, pp. 782–792. [Online]. Available: http://aclweb.org/anthology/D/D11/D11-1072.pdf

References (4)

[15] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, “Collective annotation of Wikipedia entities in Web text,” in *SIGKDD Conference*, 2009, pp. 457–466. [Online]. Available: http://www.cse.iitb.ac.in/~soumen/doc/CSAW/

[16] A. Globerson, N. Lazic, S. Chakrabarti, A. Subramanya, M. Ringgaard, and F. Pereira, “Collective entity resolution with multi-focal attention,” in *ACL Conference*, 2016, pp. 621–631. [Online]. Available: https://www.aclweb.org/anthology/P/P16/P16-1059.pdf

[17] I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji, “Joint learning of the embedding of words and entities for named entity disambiguation,” *arXiv preprint arXiv:1601.01343*, 2016. [Online]. Available: https://arxiv.org/pdf/1601.01343.pdf

[18] O.-E. Ganea and T. Hofmann, “Deep joint entity disambiguation with local neural attention,” *arXiv preprint arXiv:1704.04920*, 2017. [Online]. Available: https://arxiv.org/pdf/1704.04920.pdf

References (5)

[19] N. Lazic, A. Subramanya, M. Ringgaard, and F. Pereira, "Plato: A selective context model for entity resolution," *TACL*, vol. 3, pp. 503–515, 2015. [Online]. Available: http://anthology.aclweb.org/Q/Q15/Q15-1036.pdf

[20] N. Ge, J. Hale, and E. Charniak, "A statistical approach to anaphora resolution," in *Proceedings of the sixth workshop on very large corpora*, vol. 71, 1998, p. 76. [Online]. Available: http://www.aclweb.org/anthology/W98-1119

[21] M. Charikar, V. Guruswami, and A. Wirth, "Clustering with qualitative information," in *FOCS Conference*, 2003, pp. 524–533. [Online]. Available: http://www.cs.mu.oz.au/~awirth/pubs/awirthFocs03.pdf

[22] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in *SIGKDD Conference*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 269–278. [Online]. Available: http://www.cse.iitb.ac.in/~sunita/papers/kdd02.pdf

References (6)

[23] N. Bansal, A. Blum, and S. Chawla, “Correlation clustering,” in *FOCS Conference*, 2002, p. 238. [Online]. Available: http://www.cs.cmu.edu/~shuchi/papers/clusteringfull.pdf

[24] A. McCallum and B. Wellner, “Conditional models of identity uncertainty with application to noun coreference,” in *NIPS Conference*, 2004, pp. 905–912. [Online]. Available: https://papers.nips.cc/paper/2557-conditional-models-of-identity-uncertainty-with-application-to-noun-coreference.pdf

[25] P. Singla and P. Domingos, “Object identification with attribute-mediated dependences,” in *PKDD Conference*, Porto, Portugal, 2005, pp. 297–308. [Online]. Available: http://www.cs.washington.edu/homes/parag/papers/object-mediated-pkdd05.pdf

[26] V. Kolmogorov and R. Zabih, “What energy functions can be minimized via graph cuts?” *IEEE PAMI*, vol. 26, no. 2, pp. 147–159, Feb. 2004. [Online]. Available: http://www.cs.cornell.edu/rdz/Papers/KZ-ECCV02-graphcuts.pdf

References (7)

[27] D. M. Greig, B. T. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society*, vol. B, no. 51, pp. 271–279, 1989. [Online]. Available: http://jstor.org/stable/2345609

[28] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *International Conference on Computational Linguistics*, vol. 14, 1992, pp. 539–545. [Online]. Available: http://www.aclweb.org/website/old_anthology/C/C92/C92-2082.pdf

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

[30] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the Web," in *IJCAI*, M. M. Veloso, Ed., 2007, pp. 2670–2676. [Online]. Available: http://www.ijcai.org/papers07/Papers/IJCAI07-429.pdf

References (8)

[31] H. Poon and P. Domingos, "Unsupervised semantic parsing," in *EMNLP Conference*, 2009, pp. 1–10. [Online]. Available: http://anthology.aclweb.org/D/D09/D09-1001.pdf

[32] L. Yao, A. Haghighi, S. Riedel, and A. McCallum, "Structured relation discovery using generative models," in *EMNLP Conference*, 2011, pp. 1456–1466. [Online]. Available: http://anthology.aclweb.org/D/D11/D11-1135.pdf

[33] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *NAACL Conference*, 2013, pp. 74–84. [Online]. Available: http://www.anthology.aclweb.org/N/N13/N13-1008.pdf

[34] S. Brin, "Extracting patterns and relations from the World Wide Web," in *WebDB Workshop*, ser. LNCS, P. Atzeni, A. O. Mendelzon, and G. Mecca, Eds., vol. 1590. Valencia, Spain: Springer, Mar. 1998, pp. 172–183. [Online]. Available: http://ilpubs.stanford.edu:8090/421/1/1999-65.pdf

References (9)

[35] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *ICDL*, 2000, pp. 85–94. [Online]. Available: http://www.academia.edu/download/31007490/cucs-033-99.pdf

[36] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in *EMNLP Conference*. ACL, 2005, pp. 724–731. [Online]. Available: http://acl.ldc.upenn.edu/H/H05/H05-1091.pdf

[37] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *EMNLP Conference*, 2012, pp. 455–465. [Online]. Available: http://anthology.aclweb.org/D/D12/D12-1042.pdf

[38] G. Angeli, J. Tibshirani, J. Wu, and C. D. Manning, "Combining distant and partial supervision for relation extraction." in *EMNLP Conference*, 2014, pp. 1556–1567. [Online]. Available: http://www.anthology.aclweb.org/D/D14/D14-1164.pdf

References (10)

[39] 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*, 2011, pp. 541–550. [Online]. Available: http://anthology.aclweb.org/P/P11/P11-1055.pdf

[40] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base." in *NAACL Conference*, 2013, pp. 777–782. [Online]. Available: http://www.anthology.aclweb.org/N/N13/N13-1095.pdf

[41] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," in *AAAI Conference*, 2011, pp. 301–306. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3659/3898

[42] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *NIPS Conference*, 2013, pp. 2787–2795. [Online]. Available: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf

References (11)

[43] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix." in *ACL Conference*, 2015, pp. 687–696. [Online]. Available: http://www.aclweb.org/anthology/P/P15/P15-1067.pdf

[44] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," *arXiv preprint arXiv:1511.06361*, 2015. [Online]. Available: https://arxiv.org/pdf/1511.06361

[45] 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*, 2015, pp. 1499–1509. [Online]. Available: https://www.aclweb.org/anthology/D/D15/D15-1174.pdf

[46] P. D. Turney, "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL," in *ECML*, 2001.

References (12)

[47] J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen, "Dynamic hierarchical Markov random fields and their application to Web data extraction," in *ICML*, 2007, pp. 1175–1182. [Online]. Available: http://www.machinelearning.org/proceedings/icml2007/papers/215.pdf

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

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

[50] V. Hristidis, L. Gravano, and Y. Papakonstantinou, "Efficient IR-style keyword search over relational databases," in *VLDB Conference*, 2003, pp. 850–861. [Online]. Available: http://www.db.ucsd.edu/publications/VLDB2003cr.pdf

[51] G. Jeh and J. Widom, "Scaling personalized web search," in *WWW Conference*, 2003, pp. 271–279. [Online]. Available: http://www2003.org/cdrom/papers/refereed/p185/html/p185-jeh.html

References (13)

[52] T. H. Haveliwala, "Topic-sensitive PageRank," in *WWW Conference*, 2002, pp. 517–526. [Online]. Available: http://www2002.org/CDROM/refereed/127/index.html

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

[54] M. J. Cafarella, C. Re, D. Suciu, O. Etzioni, and M. Banko, "Structured querying of web text: A technical challenge," in *CIDR*, 2007, pp. 225–234. [Online]. Available: http://www-db.cs.wisc.edu/cidr/cidr2007/papers/cidr07p25.pdf

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

References (14)

[56] T. Cheng, X. Yan, and K. C.-C. Chang, "EntityRank: Searching entities directly and holistically," in *VLDB Conference*, Sep. 2007, pp. 387–398. [Online]. Available: http://www-forward.cs.uiuc.edu/pubs/2007/entityrank-vldb07-cyc-jul07.pdf

[57] S. Chakrabarti, "Dynamic personalized PageRank in entity-relation graphs," in *WWW Conference*, Banff, May 2007. [Online]. Available: http://www.cse.iitb.ac.in/~soumen/doc/netrank/

[58] P. Sarkar, A. W. Moore, and A. Prakash, "Fast incremental proximity search in large graphs," in *ICML*, 2008, pp. 896–903. [Online]. Available: http://icml2008.cs.helsinki.fi/papers/565.pdf

[59] D. Milne and I. H. Witten, "Learning to link with Wikipedia," in *CIKM*, 2008, pp. 509–518. [Online]. Available: http://www.cs.waikato.ac.nz/~dnk2/publications/CIKM08-LearningToLinkWithWikipedia.pdf

References (15)

[60] G. Kasneci, F. M. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath, and G. Weikum, “NAGA: harvesting, searching and ranking knowledge,” in *SIGMOD Conference*. ACM, 2008, pp. 1285–1288. [Online]. Available: http://www.mpi-inf.mpg.de/~kasneci/naga/

[61] F. M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A core of semantic knowledge unifying WordNet and Wikipedia,” in *WWW Conference*. ACM Press, 2007, pp. 697–706. [Online]. Available: http://www2007.org/papers/paper391.pdf

[62] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS Conference*, 2013, pp. 3111–3119. [Online]. Available: https://goo.gl/x3DTzS

[63] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation.” in *EMNLP Conference*, vol. 14, 2014, pp. 1532–1543. [Online]. Available: http://www.emnlp2014.org/papers/pdf/EMNLP2014162.pdf

References (16)

[64] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine Learning*, vol. 81, no. 1, pp. 53–67, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10994-010-5205-8

[65] S. Sarawagi, "Information extraction," *FnT Databases*, vol. 1, no. 3, 2008. [Online]. Available: http://www.cse.iitb.ac.in/~sunita/papers/ieSurvey.pdf