Organizing Web Information

CS 728

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

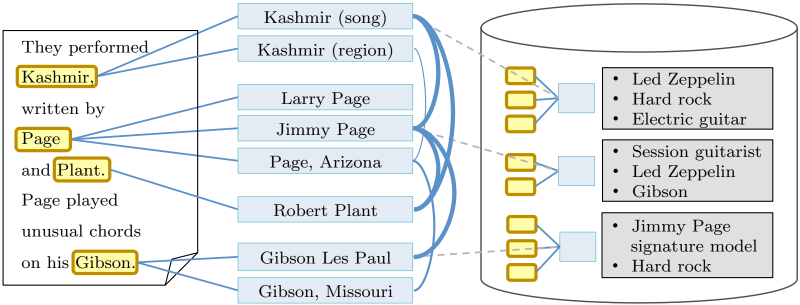
Named entity disambiguation

Entity linking — working definition and motivation

- From coarse NER to fine types to specific entities with canonical IDs in knowledge graph (KG), e.g., `http://en.wikipedia.org/wiki/Michael_Jordan` or `http://en.wikipedia.org/wiki/Michael_I._Jordan`
- Many choices of KGs: Wikipedia, WikiData, Freebase, Google KG, Bing Satori, . . .
- An entity can have many aliases: Michael Jordan, Mike, Jordan
- Conversely, *Jordan* can refer to river, country, and lots of people
- A passage may mention[6] an entity; around the mention m is a context c from which we can observe context features
- If the mention string matches an alias of an entity e, the entity becomes a candidate
- $\Gamma(m)$ is the set of all candidates of mention m

Entity linking — working definition and motivation (2)

- For each mention, the goal is to choose one or zero (out-of-KG, reject, null, $\perp$, $\textsc{na}$) candidate
- Each mention is a multiclass, single-label classification problem, but they are inter-related



- Entity label at mention m_i is y_i; gold label is y_i^*

Entity linking — working definition and motivation (3)

- Motivation: complex query responses involving joins
  - Company `?c` in Korea makes phone `?p` under \$400 with OLED display — instantiate all possible $\langle \texttt{?c}, \texttt{?p} \rangle$
  - Need to recognize `?c`, `?p` as (single) company and phone in different contexts provide evidence for subqueries

[6] Detecting mention boundaries is difficult [10] but is outside our scope.

Some distinctions from WSD

- Word sense disambiguation (WSD) is largely about common words, not references to specific entities
  - 42 senses of "run" in WordNet
  - Part of speech helps a fair bit
  - Identifying mention boundary is easy
- Entity catalog typically richer info source than dictionary
  - Broader category system
  - Part of speech is largely "proper noun" and not as helpful
- Entity disambiguation goals:
  - Identify that a sequence of tokens is a potential mention
  - Capture suitable context around to form spot s
  - Assign s to a suitable entity γ in catalog
  - Or claim that there is no suitable γ

Why annotate?

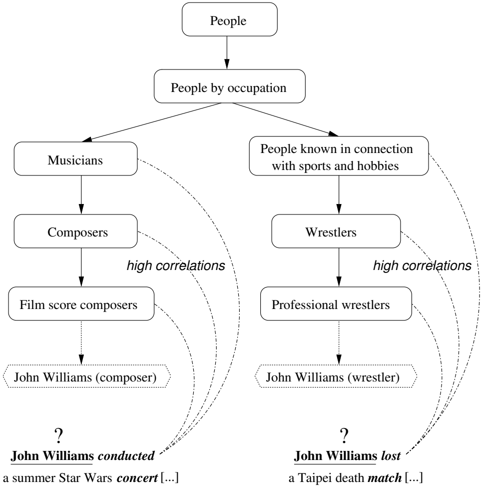
- Make raw text look like Wikipedia with definitional and informational links (most systems)
  - Annotate first occurrence only
  - Annotate only on-topic entities
  - Use discretion to avoid "hyperlink fatigue"
- Index the annotations to enable advanced search (our focus)
  - Exhaustive annotation
  - Make no whole-document topic judgment

More about catalog representation

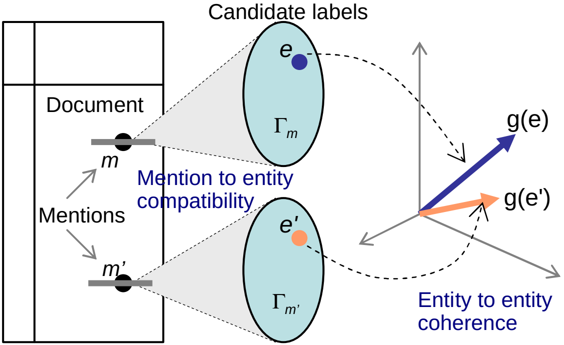
- Pattern after WordNet, Wikipedia, Freebase, ...
- Each entity γ has associated description or definition page
- Descriptions link to other related entities γ'
- Entities belong to one or more categories
- Categories (physicist) are subcategories of others (scientist)
- Links may be "incidental"
- Categories and super-categories may be noisy: *Machine learning researcher* more meaningful than *Living people* or *Year of birth missing*
- Cycles in is-a "hierarchy"?

Local signals to choose e from $\Gamma(m)$

- Match between context and entity
- Entity representations
  - Text on definition pages in Wikipedia
  - Text from gold mention contexts
  - Types that contain the entity
- Between context and types containing entity [11]
- Between page topic/s and entity type/s [12]



Integrating local and global signals [13, 12]



- Some entity pairs are more coherent than others
- Coherence may be measured in different ways
- Choose per-mention entity labels to maximize pairwise coherence as well as local compatibility
- Intractable in general; heuristic approximations common

SemTag [14]

- Used Stanford TAP ontology (72,000 entities)
- Set of classes C, subclass relation $S \subseteq C \times C$, set of instances (entities) I, many-to-many type relation $T \subseteq I \times C$
- i has class c_1 and c_1 subclass of c_2 implies i has class c_2
- Entity taxonomy is a DAG, $\pi(v)$ is the path up from v to root node r
- Taxonomy node v has label set $L(v)$, e.g., nodes corresponding to cats, football, computers and cars all contain the label 'jaguar'

SemTag output example

```
The <resource ref="http://tap.stanford.edu/BasketballTeam
_Bulls">Chicago Bulls</resource> announced yesterday that
<resource ref="http://tap.stanford.edu/AthleteJordan,_Mich
ael">Michael Jordan</resource> will ...
```

- Functionally identical to inserting Wikipedia links in free-form text
- Wikipedia is more organic than TAP; has poorer quality category hierarchy

SemTag disambiguation

- $\mathrm{sim}(u, s) \in [0, 1]$ is a local similarity between catalog node u and (context of) spot s
- $\mathrm{sim}(\cdot, \cdot) = \frac{1}{2}$ is "most uncertain"
- Node v is eligible for spot s if

  $$\text{root } r \neq \arg\max_{u \in \pi(v)} \mathrm{sim}(u, s)$$

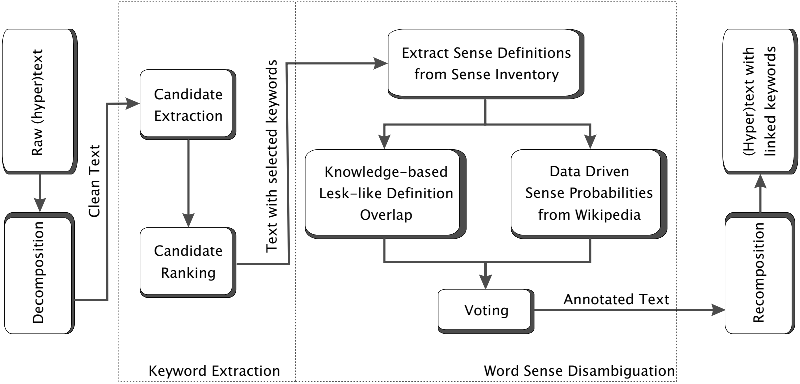
  i.e., some node on $\pi(v)$ other than root most similar to s
- Supplement eligibility with human-judged scores of reliability at each node u
  - m_u^a = probability that spots for subtree rooted at u are "on topic"
  - m_u^s = probability that automatic eligibility judgment is correct

SemTag TBD algorithm

- To decide whether to link spot s to node v ...
- Find nearest ancestor u of v that has human-judged reliability scores
- If $|\frac{1}{2} - m_u^a| > |\frac{1}{2} - m_u^s|$, return $\mathrm{sign}(m_u^a - \frac{1}{2})$
- Else if $m_u^s > \frac{1}{2}$ (eligibility judgment is often correct), return $\mathrm{eligible}(c, u)$
- Else (eligibility judgment is often wrong) return $1 - \mathrm{eligible}(c, u)$

(Can regard as a simple hand-tuned form of stacked learning)

Wikify! [15]



- Two-phase process
- First identify token spans “worthy of annotation”
- Then choose entity labels

Sample annotations

In 1834, Sumner was admitted to the [[bar (law)|bar]] at the age of twenty-thre, and entered private practice in Boston.

It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every [[bar (music)|bar]].

Vehicles of this type may contain expensive audio players, televisions, video players, and [[bar (counter)|bar]]s, often with refrigerators.

Jenga is a popular beer in the [[bar (establishment)|bar]]s of Thailand

---

This is a disturbance on the water surface of a river or estuary, often cause by the presence of a [[bar (landform)|bar]] or dune on the riverbed.

Choosing token spans to annotate ("spotting")

- Wikify! follows the Wikipedia philosophy
- Use some score to rank candidate spans
- TFIDF of a token in a document
- χ^2 test:

count of token in doc	count of all other tokens in doc
count of token in other docs	count of all other tokens in other docs

- "Keyphraseness" — In how many Wikipedia documents is the same term made a link anchor?
- (They only consider as candidates words which appear at least five times in Wikipedia)

Disambiguation

Wikify! compares two local techniques:

- "Knowledge-based approach" — similarity between Wikipedia page text of entity γ and context words in spot s
- "Data-driven approach" — similarity between context of known links to γ and context words in spot s
- "Context" consists of ± 3 words around mention, their parts of speech, salient words chosen from whole document

Results

- "Data-driven" better than "knowledge-based"
- Consensus (agreement) has highest precision

	Words		Evaluation		
Method	(A)	(C)	(P)	(R)	(F)
			Baselines		
Random baseline	6,517	4,161	63.84	56.90	60.17
Most frequent sense	6,517	5,672	87.03	77.57	82.02
			Word sense disambiguation methods		
Knowledge-based	6,517	5,255	80.63	71.86	75.99
Feature-based learning	6,517	6,055	92.91	**83.10**	**87.73**
Combined	5,433	5,125	**94.33**	70.51	80.69


Welcome to Wikify! - Mozilla Firefox
Wiki fier
Browse...
http://news.bbc.co.uk/2/hi/south_asia/539
Wikify!
HOME | HELP | ABOUT | CONTACT
UK version International version | About the versions
BBC NEWS
BBC News in video and audio
Last Updated: Thursday, 28 September 2006, 19:20 GMT
E-mail this to a friend
Printable version
Nato to extend Afghan operations
Nato has announced that it will extend its mission in Afghanistan to cover the whole of the insurgency-hit country.
The move will take the alliance into the eastern parts of Afghanistan and bring up to 12,000 American troops under Nato command.
Nato will now command more than 30,000 troops in Afghanistan
A Nato official said the decision would be implemented in the next few weeks.
The announcement came as the US military said that militant attacks near the Pakistani border had tripled in some areas.
The rise in activity comes despite a peace agreement meant to end violence by pro-Taleban militants in Pakistan's North Waziristan border area.
News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Health
Science/Nature
Technology
Entertainment
Have Your Say
In Pictures
Country Profiles
Special Reports
Programmes
RELATED BBC SITES
Afghanistan
From Wikipedia, the free encyclopedia
Military of the United States
From Wikipedia, the free encyclopedia
The military of the United States, officially known as the United States Armed Forces, consist of the
United States Army
United States Marine Corps
United States Navy
United States Air Force
United States Coast Guard
Afghanistan at-a-glance
Done

Modeling local compatibility

- Feature vector $f_s(\gamma) \in \mathbb{R}^d$ expresses local textual compatibility between (context of) spot s and candidate label γ
- One element of $f_s(\gamma)$ might be the TFIDF cosine similarity between tokens from the context of spot s (say ± 10 tokens) and whole page of description for entity γ
- Another element may be derived of "anchor text" match:
  - Find all links to γ from within Wikipedia
  - Collect anchor text from all these links in a bag of words
  - Find TFIDF cosine similarity between this bag and the spot context s

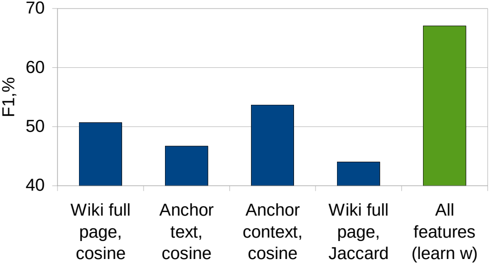
The sense probability prior

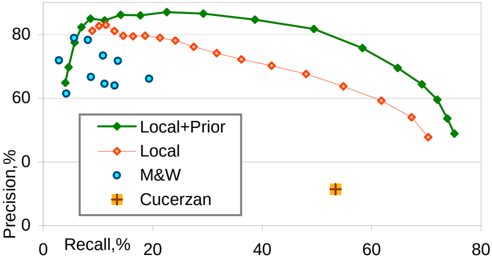
- What entity does "Intel" refer to?
  - Chip design and manufacturing company
  - Fictional cartel in a 1961 BBC TV serial
- $\Pr_0(\gamma|s)$ is very high for chip maker, low for cartel
- Append element $\log \Pr_0(\gamma|s)$ to $f_s(\gamma)$
- "$\log$" will be explained later

Node score

- Node scoring model $w \in \mathbb{R}^d$
- Node score defined as $w^\top f_s(\gamma)$
- w is trained to give suitable relative weights to different compatibility measures and aggregate the evidence
- During test time, greedy choice local to s would be $\arg\max_{\gamma \in \Gamma_s} w^\top f_s(\gamma)$
- Early algorithms are variations on this theme

Effect of learning single-mention scores



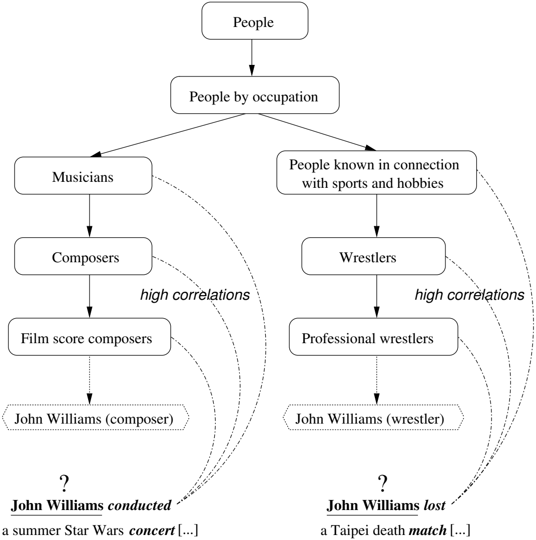


- Learning w is better than commonly-used single features
- Enough to beat some collective approaches (soon)

Limitations of $\mathrm{sim}(\gamma, s)$

- Training data is sparse
- Direct overlap of words between description of entity γ and context of spot s may be limited
- But overlap between ancestors of γ and context of s may be more reliable

Word-category correlations



Designing tree kernels

- Let $C(\gamma)$ be all ancestor categories of entity γ
- Let $T(s)$ be the text in the context of spot s
- For every word w and every all categories c, define a feature

$$\phi_{w,c}(s,\gamma) = \begin{cases} 1 & \text{if } w \in T(s) \text{ and } c \in C(\gamma) \\ 0 & \text{otherwise} \end{cases}$$

- Run through all possible w, c, e.g., (“conducted”, $musician$), (“concert”, $wrestler$)
- Pad $(\phi_{w,c})$ with local compatibility features
- Finally, get feature vector $\Phi(s,\gamma)$

Learning

- Model as classification: correct/incorrect (s, γ) pair should be labeled $+1/-1$ respectively
- Similar to sequence labeling: $\arg\max_\gamma w^\top \Phi(s, \gamma)$; same max-margin training
- What about spots that do not have any suitable entity in the catalog?
- Out-of-catalog entity $\hat{\gamma}$, with $C(\hat{\gamma}) = \varnothing$ and $T(\hat{\gamma}) = \varnothing$
- One last feature element $\phi_{\hat{}}(s, \gamma) = [\![\gamma = \hat{\gamma}]\!]$
- Equivalent to automatically learning a (lower) threshold on $w^\top \Phi(s, \gamma)$

Tree kernel results

Data set	TreeKernel	TextOnly
People by occupation, top 110	0.772	0.615
Ditto, all 540	0.684	0.558
Ditto, categories with ≥ 20 entities	0.680	0.554

- Summary: tree kernel better than comparing only text

Modeling entity relatedness from catalog

- Some entity pairs are more compatible than others
- Better to choose per-mention entity labels to maximize pairwise compatibility
- Compatibility may have different notions
- Entities belong to related types, e.g., soccer coaches, clubs, players [13, 12]
- Frequently co-cited from Web/Wikipedia pages [16]
- Entities connected by short path in knowledge graph [17]
- (Similarity between vector embeddings of entities based on corpus mentions — soon)
- How related are two entities γ, γ' in Wikipedia?
- Embed γ in some space using $g : \Gamma \to \mathbb{R}^c$
- Define relatedness $r(\gamma, \gamma') = g(\gamma) \cdot g(\gamma')$ or related

Modeling entity relatedness from catalog (2)

- Cucerzan's proposal: $c =$ number of categories; $g(\gamma)[\tau] = 1$ if γ belongs to category τ, 0 otherwise

$$r(\gamma, \gamma') = \frac{g(\gamma)^\top g(\gamma')}{\sqrt{g(\gamma)^\top g(\gamma)}\sqrt{g(\gamma')^\top g(\gamma')}},$$

(standard cosine)

- Milne and Witten's proposal: $c =$ number of Wikipedia pages; $g(\gamma)[p] = 1$ if page p links to page γ, 0 otherwise

$$r(\gamma, \gamma') = \frac{\log \frac{|g(\gamma) \cap g(\gamma')|}{|g(\gamma) \cup g(\gamma')|}}{\log \frac{c}{\min\{|g(\gamma)|, |g(\gamma')|\}}}$$

- Related to Jaccard
- With voice of small inlink sets attenuated
- Combination of above [18]

A joint local+global objective

- Notation: mentions written variously as m_i, s_i; s_i includes m_i and features from context c_i
- Entity labels written variously as γ_i, y_i, e_i
- $\Psi(e_i, m_i, c_i)$ is the local score of entity e_i for mention m_i with context c_i
- $\Phi(e_i, e_j)$ is the pairwise coherence between the entities chosen for mentions i, j
- For whole document, let $\boldsymbol{e}, \boldsymbol{m}, \boldsymbol{c}$ be the sequence of n entity labels, mentions, and contexts
- Overall objective is to maximize wrt $\boldsymbol{e}$

$$g(\boldsymbol{e}, \boldsymbol{m}, \boldsymbol{c}) = \underbrace{\frac{1}{n} \sum_i \Psi(e_i, m_i, c_i)}_{\text{local}} + \underbrace{\frac{1}{\binom{n}{2}} \sum_{i \neq j} \Phi(e_i, e_j)}_{\text{global}}$$

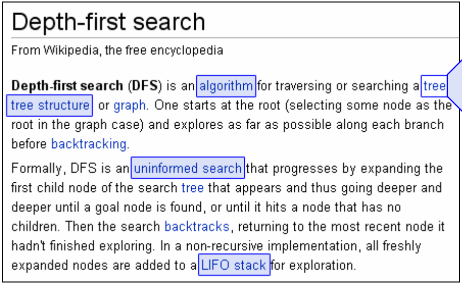
A joint local+global objective (2)

- (Conditional) probabilistic graphical model with complete graph
- Aka the quadratic assignment problem
- Difficult NP-hard problem
- Heuristics: leave-one-out [19], easy-mention-first [16], hill-climbing [13, 20], LP relaxation [13], multifocal attention [21]

Leave-one-out disambiguation [19]

- Let $\Gamma_0 = \bigcup_i \Gamma(m_i)$ be all possible entity disambiguations for all mentions on a page
- Precompute the average entity representation vector $g(\Gamma_0) = \sum_{\gamma \in \Gamma_0} g(\gamma)$
- Score of candidate label γ for spot s depends on two factors multiplied together
- The local compatibility score as before
- $g(\gamma)^\top g(\Gamma_0 \setminus \{\gamma\}) = g(\gamma)^\top \sum_{\gamma' \in \Gamma_0 \setminus \gamma} g(\gamma')$
- Note that $\Gamma_0 \setminus \gamma$ still contains contributions from entities that cannot be used simultaneously to label the page
- $g(\Gamma_0 \setminus \gamma)$ may not be a representative feature vector

Commonness, usefulness, relatedness



Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree, tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

- "Tree" has many senses, common and rare
- But a low probability sense may be the correct one, based on relatedness to unambiguous anchor entities mentioned near "tree"
- Not all anchors equally useful: "until" vs. "LIFO"

Milne and Witten's recipe

- Identify unambiguous spots $S_!$ from all spots S_0
- Denote $\Gamma_! = \bigcup_{s \in S_!} \Gamma_s$, note that $\Gamma_! \xleftrightarrow{1:1} S_!$
- Ambiguous spot $s \mapsto \Gamma_s$, have to pick $\gamma \in \Gamma_s$
- Each candidate γ is scored based on three signals

  Commonness of γ, i.e., sense probability prior $\Pr_0(\gamma|s)$

  Average relatedness to anchor entities $\gamma_!$, weighted by the usefulness $u(\gamma_!)$ of $\gamma_!$

$$\frac{\sum_{\gamma_! \in \Gamma_! \setminus \gamma} u(\gamma_!) r(\gamma, \gamma_!)}{\sum_{\gamma_! \in \Gamma_! \setminus \gamma} u(\gamma_!)}$$

$$\text{where } u(\gamma) = \sum_{\gamma'' \in \Gamma_! \setminus \gamma'} r(\gamma', \gamma'')$$

  Overall context quality for the spot, $\sum_{\gamma_!} u(\gamma_!)$

Milne and Witten's recipe (2)

- These three signals are presented as features to a classifier (bagged decision tree worked best)
- The label is whether γ is correct for s

M&W results

	recall	precision	f-measure
Random sense	56.4	50.2	53.1
Most common sense	92.2	89.3	90.7
Medelyan *et al.* (2008)	92.3	93.3	92.9
Most valid sense	95.7	**98.4**	**97.1**
All valid senses	**96.6**	97.0	96.8

- Random sense gives precision over $\frac{1}{2}$, only around two senses per spot
- Recall is as per (reticent) Wikipedia annotation policy

correct	76.4
incorrect (wrong destination)	0.9
incorrect (irrelevant and/or unhelpful)	19.8
incorrect (unknown reason)	2.9

Hill-climbing [20]

- Two stages, ranker followed by linker
- Ranker obtains best non-null label for each mention
- Linker decides whether to replace best label with NA

for each mention m_i **do**
    construct disambiguation candidates Γ_i
    run **ranker** to get best non-null disambiguation y_i
for mentions m_i in some arbitrary order **do**
    if changing y_i to null improves collective objective **then**
        commit to change

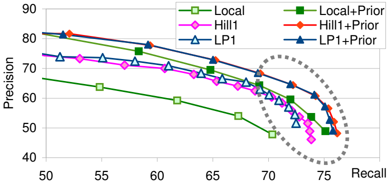
More details

Integer program

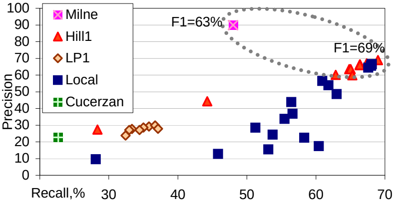
- Let i index mentions and e index candidate entities
- Decision variable $z_{ie} \in \{0, 1\}$ is 1 if mention i gets label e and 0 otherwise
- For each mention i, $\sum_{e \in \Gamma_i} z_{ie} \leq 1$ (zero or one label per mention from among candidates)
- Local node log-potential for mention i is $\phi_i(e)$
- Local objective is $\sum_i \sum_e \phi_i(e) z_{ie}$
- Auxiliary decision variables $p_{i,e,i',e'} \in \{0, 1\}$ for all mention and label pairs
- Constraints for all i, e, i', e', $p_{i,e,i',e'} \leq z_{ie}$ and $p_{i,e,i',e'} \leq z_{i'e'}$
- Global objective is $\sum_{i,e,i',e'} p_{i,e,i',e'} \psi_{ii'}(e, e')$
- Relax to $z_{i,e}, p_{i,e,i',e'} \in [0, 1]$ (not a nice relaxation, cannot round to provably good discrete solutions)

Benefits of collective labeling

- Two different data sets (Web, newswire)
- Can significantly push recall while preserving precision
- Improves upon Milne&Witten [16], Cucerzan [19]



Web pages



News articles

Multifocal attention [21]

- Consider again the all-pairs global term $\sum_{i \neq j} \Phi(e_i, e_j)$
- Entities in doc may not all be in one type cluster; e.g., e_i may be a politician and e_j a real-estate baron
- KG may not know of common type-to-type relations, e.g., cricketers and business tycoons, or politicians and real estate barons
- Less salient entity e_i may not find enough Φ support from all other entities e_j
- Asserting all-pairs potentials across coherent clusters needlessly adds noise floor to objective
- Discussed by Kulkarni et al. [13] but not addressed

Single link baseline

- As an extreme simplication of the clique potential, for each mention, find one best supporter

$$g_{\mathsf{SL}}(\boldsymbol{y}) = \prod_i s_i(y_i) \left[\max_j s_{ij}(y_i, y_j) \right]$$

- $\boldsymbol{y}$ is the vector of entity labels assigned to all mentions in a document
- $s_i(y_i)$ is the local score for entity label y_i for mention/spot i
- MAP inference is still intractable
  - If j is the best supporter of i, is i necessarily the best supporter of j?
- Approximate by message passing (loopy belief propagation) on factor graph
- Factor a_i for each mention i

Single link baseline (2)

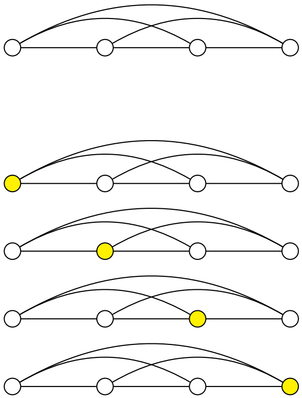
- Each factor connects to all (mention) nodes, but best supporter makes message passing practical
- Message from a_k to mention i is

$$n_{a_k \to i}(y_i) = \max_{\boldsymbol{y}_{\setminus i}} \Big[\psi_k(y_i, \boldsymbol{y}_{\setminus i}) \prod_{j \neq i} m_{j \to a_k}(y_j) \Big]$$

- Belief in $\boldsymbol{y}$ based on incoming messages from all factors

Relaxing to a star model

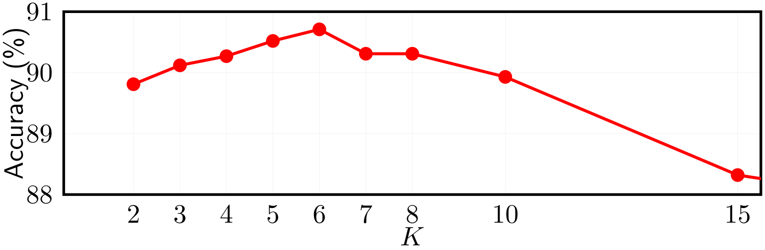
- Give up global consistency for tractability
- In turn, make each mention center of a star
- Assign label to each spoke separately to maximize support for hub
- Support for label y_i from mention j is $q_{ij}(y_i) = \max_{y_j}[s_{ij}(y_i, y_j) + s_j(y_j)]$
- Score function for mention i is $f_i(y_i) = s_i(y_i) + \sum_{\text{all } j \neq i} q_{ij}(y_i)$
- Predict y_i by maximizing above score
- Next step: replace all $j \neq i$ with something more robust
- In what follows, let $\boldsymbol{q}_i(y_i) = \langle q_{i1}(y_i), \ldots, q_{in}(y_i) \rangle$ be the sequence of support from other mentions to mention i



You need only six good friends

- Star model with top-K supporters:
  - When choosing e_i, set other e_j to get the best top-K supporters e_j, rather than all $n-1$
  - Later, when setting e_j, do not constrain e_i to be the label earlier chosen
- Best support for label e_i from mention j is $q_{ij}(e_i) = \max_{e_j} [\Psi(e_j) + \Phi(e_i, e_j)]$
- Star model with all $n-1$ supporters amounts to overall score $f_i(e_i) = \Psi(e_i) + \sum_{j \neq i} q_{ij}(e_i)$
- Let $\boldsymbol{q}_i(e_i) = \langle q_{i1}(e_i), \ldots, q_{in}(e_i) \rangle$ be the sequence of supports from other mentions to mention i
- Given support sequence $\boldsymbol{q}$, let $\text{amx}_K(\boldsymbol{q})$ be the sum of the largest K elements of $\boldsymbol{q}$
- Redefine score function for ith mention as $f_i(e_i) = \Psi(e_i) + \text{amx}_K(\boldsymbol{q}_i(e_i))$

You need only six good friends (2)



- Plot accuracy against K
- Single supporter too little to go by
- All $n-1$ supporters too much to ask for
- Clear peak at $K=6$
- K supporters get full backprop, others get none
- From K-max to soft-K-max

Multifocal last step: from max to soft-max

- Find maximum element in non-negative vector q is equivalent to $\max_{u \in \Delta} u \cdot q$
- Δ is the unit simplex
- u will concentrate on one corner of Δ
- Anneal with entropy: $\max_{u \in \Delta} u \cdot q + H(u)/\beta$
- Easy to see solution as $u_i \propto \exp(\beta q_i)$
- In other words, adding entropic annealing to max gives us soft-max
- In standard multiclass classification, benefit of soft-max is continuous differentiability
- Can backprop to downstream model components

Soft multifocal attention

- Recall $\boldsymbol{q} = \langle q_{i1}(y_i), \ldots, q_{in}(y_i) \rangle$ is the vector of supports for y_i
- Add entropy term to **amx** to get **smx**:

$$\mathsf{smx}_K(q) = \max_{u \in \Delta_K} \left[q \cdot u - \frac{1}{\beta} \sum_i u_i \log u_i \right]$$

- Here Δ_K is the K-simplex: $u \geq \vec{0}$ and $\|u\|_1 = K$
- smx_K can be computed easily and is differentiable

HW Apply to fine typing and other applications where softmax gives excessively skewed attention

$$\begin{tikzcd}
\mathsf{max} \arrow[r] \arrow[d] & \mathsf{softmax} \arrow[d] \\
\mathsf{amx}_K \arrow[r] & \mathsf{smx}_K
\end{tikzcd}$$

Soft multifocal attention

- Note $\text{amx}_K(\boldsymbol{q}) = \max_{\vec{0}\le\boldsymbol{z}\le\vec{1}} \boldsymbol{z}\cdot\boldsymbol{q}$ s.t. $\sum_j z_j = K$
- Replace amx_K with soft K-max
  $\text{smx}_K(\boldsymbol{q}) = \max_{\vec{0}\le\boldsymbol{z}\le\vec{1}} \boldsymbol{z}\cdot\boldsymbol{q} \underbrace{-\sum_j z_j \log z_j}_{\text{entropy}}$ s.t. $\sum_j z_j = K$
- Generalizes softmax
- Used to train model weights inside Ψ, Φ

System	Alias-entity map	Accuracy%
Lazic+ 2015	Older KG	86.4
Our baseline	Latest KG	87.9
Single link	Latest KG	88.2
Multifocal	Latest KG	89.5
Chisholm+ 2015	YAGO	88.7
Our baseline	YAGO+KG	85.2
Single link	YAGO+KG	86.6
Multifocal	YAGO+KG	91.0
Multifocal	KG+HP	92.7

Soft multifocal attention (2)

- Within each alias-entity map, single-link and multifocal are the best
- Baseline and single link degrade when alias map changes from KG to YAGO+KG (larger ambiguity), but multifocal improves
- Similar consistent gains in TAC 2010, 2011, 2012
- What's missing? Entity embeddings

Using entity embeddings [22]

- Three-part optimization
- Overall likelihood fitted through simultaneous maximization
  $\mathcal{L} = \mathcal{L}_w + \mathcal{L}_e + \mathcal{L}_a$
  Word-word: $\mathcal{L}_w$, standard word2vec on text corpus
  Entity-entity: $\mathcal{L}_e$, as expressed through KG
  Word-entity: $\mathcal{L}_a$, connecting mention context words and entity embeddings
- e, e' are related if there is a link between them in the KG, and $e \neq e'$, in which case we want large

$$\mathcal{L}_e = \sum_{e,e'} \log \Pr(e'|e), \qquad \text{where}$$

$$\Pr(e'|e) = \frac{\exp(\boldsymbol{u}_e \cdot \boldsymbol{v}_{e'})}{\sum_e \exp(\boldsymbol{u}_e \cdot \boldsymbol{v}_e)}$$

Using entity embeddings [22] (2)

- As in skip-gram, predict mention context words given focus entity ID
- Let M_e be mentions of entity e, $m \in M_e$ be one mention, and $w \in m$ a mention word

$$\mathcal{L}_a = \sum_e \sum_{m \in M_e} \sum_{w \in m} \log \Pr(w|e),$$

where
$$\Pr(w|e) = \frac{\exp(\boldsymbol{v}_w \cdot \boldsymbol{u}_e)}{\sum_{w'} \exp(\boldsymbol{v}_{w'} \cdot \boldsymbol{u}_e)}$$

- As is common, softmax is replaced by negative samples

Inference with coherence

- Given a document with many mention spots
- For each mention, compute context vector as average of neighboring word vectors
- (Nothing more fancy like convnet or RNN)
- Set initial entity labels using cosine with context vectors
- Now define the coherence of an entity with others as average cosine between entity vectors
- Reassign most coherent label in a second step
- Crude two-step loopy BP?

Joint word-entity embeddings: NED results

	CoNLL (Micro)	CoNLL (Macro)	TAC10 (Micro)
Yamada et al., 2016	**93.1**	**92.6**	**85.2**
Hoffart et al., 2011	82.5	81.7	-
He et al., 2013	85.6	84.0	81.0
Chisholm & Hachey, 2015	88.7	-	80.7
Pershina et al., 2015	91.8	89.9	-

Attention on mention context [23]

- Jointly pre-embed all words w and entities e in training corpus (Wikipedia, say) to (focus) embeddings x_w, x_e
- Given a mention m with candidates $\Gamma(m)$, mention context c mentioning entity $e \in \Gamma(m)$, for each word w in the context, compute the importance of w as

$$u(w) = \max_{e \in \Gamma(m)} x_e^\top \boldsymbol{A} x_w,$$

  where $\boldsymbol{A}$ is a global (diagonal) matrix to be trained
- Intention: $u(w)$ should be large if w is strongly associated with at least one candidate entity, otherwise small
- Sort by decreasing $u(w)$ and prune context to top-K
- Now let surviving context words compete for attention:

$$\beta(w) = \exp\big(u(w)\big) \big/ \textstyle\sum_{w'} \exp\big(u(w')\big)$$

Attention on mention context [23] (2)

- Compute similarity between x_e and x_w and add up, weighted by attention:

$$\Psi(e,c) = \textstyle\sum_w \beta(w)\, x_e^\top \boldsymbol{B} x_w\,,$$

  where $\boldsymbol{B}$ is another global diagonal matrix to be trained
- Note, very frugal model so far, only $2D$ model weights, where embeddings are in $\mathbb{R}^D$
- Finally, combine with (empirical) mention prior $\Pr(e|m)$:

$$\Psi(e,m,c) = N\big(\Psi(e,c), \log\Pr(e|m)\big),$$

  where N is a 2-layer fully-connected network with 100 hidden units and ReLU nonlinearities

Attention on mention context [23] (3)

- For training, use standard hinge loss

$$\operatorname*{argmin}_{\boldsymbol{A},\boldsymbol{B},N,\ldots} \sum_{m} \sum_{e \in \Gamma(m)} \left[\clubsuit - \Psi(e^*, m, c) + \Psi(e, m, c) \right]_+ ,$$

where $\clubsuit$ is a tuned margin

- Local attention model results:

Methods	AIDA-test-b
Mention prior $\Pr(e\|m)$	71.9
(Lazic et al., 2015)	86.4
(Yamada et al., 2016)	87.2
(Globerson et al., 2016)	87.9
Ganea+ (local, K=100, R=50)	**88.8**

- Network N benefits from nonlinearity

Document-level deep model

- Now we bring back in global coherence between entity labels
- For a whole document, let $\boldsymbol{e}, \boldsymbol{m}, \boldsymbol{c}$ be the sequence of n entity labels, mentions, and contexts
- Fully connected pairwise random field

$$g(\boldsymbol{e}, \boldsymbol{m}, \boldsymbol{c}) = \frac{1}{n} \sum_i \Psi(e_i, m_i, c_i) + \frac{1}{\binom{n}{2}} \sum_{i<j} \Phi(e_i, e_j),$$

where $\quad \Phi(e, e') = x_e^\top \boldsymbol{C} x_{e'}$

- Note, all mention pairs
- $\boldsymbol{C}$ is another diagonal weight matrix to be trained
- Inference amounts to finding $\operatorname{argmax}_{\boldsymbol{e}} g(\boldsymbol{e}, \boldsymbol{m}, \boldsymbol{c})$, given observed $\boldsymbol{m}, \boldsymbol{c}$
- Back to (max-product) message-passing

Document-level deep model (2)

- In iteration t, mention m_i votes for entity candidate $e' \in \Gamma(m_j)$ using outgoing (log) message

$$m_{i\to j}^{t+1}(e') = \max_{e\in\Gamma(m_i)} \left[\Psi(e, m_i, c_i) + \Phi(e, e') + \sum_{k\neq j} \overline{m}_{k\to i}^{t}(e)\right]$$

- The incoming messages would ordinarily be just log-beliefs:

$$\overline{m}_{i\to j}^{t}(e) = \log \mathsf{softmax}\big(m_{i\to j}^{t}(e)\big)$$

- In practice, damping with $\delta \in (0, 1]$ helps stability and convergence:

$$\overline{m}_{i\to j}^{t}(e) = \log\Big[\delta\, \mathsf{softmax}\big(m_{i\to j}^{t}(e)\big) + (1-\delta)\exp(\overline{m}_{i\to j}^{t-1}(e))\Big]$$

Document-level deep model (3)

- Unroll BP to T time steps, resulting in final beliefs

$$\mu_i(e) = \Psi(e, m_i, c_i) + \sum_{k \neq i} \overline{m}_{k \to i}^{T}(e)$$

$$\overline{\mu}_i(e) = \frac{\exp\big(\mu_i(e)\big)}{\sum_{e' \in \Gamma(m_i)} \exp\big(\mu_i(e)\big)}$$

- Given the above inference procedure, we can use it for training as well
- Given gold entity labels, express hinge loss wrt final beliefs:

$$\operatorname*{argmin}_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},N} \sum_{m} \sum_{e \in \Gamma(m)} \left[\spadesuit - \overline{\mu}_i(e^*) + \overline{\mu}_i(e)\right]_+$$

- Hinge loss assessed wrt per-variable marginals
- Everything is still end-to-end (sub)differentiable

Document-level deep model (4)

- May be simpler (but possibly less accurate) than sampling negative instances and expressing objective as hinge loss corresponding to

$$\forall \boldsymbol{e}_- : \qquad g(\boldsymbol{e}_+, \boldsymbol{m}, \boldsymbol{c}) \geq g(\boldsymbol{e}_-, \boldsymbol{m}, \boldsymbol{c}) + \spadesuit$$

Ganea et al.: global results

Global method	AIDA-test-b
(Huang et al., 2015)	86.6
(Ganea et al., 2016)	87.6
(Chisholm and Hachey, 2015)	88.7
(Guo and Barbosa, 2016)	89.0
(Globerson et al., 2016)	91.0
(Yamada et al., 2016)	91.5
Ganea+ (global)	**92.22±0.14**

- Impressive gains with very few model weights!
- Even more impressive that tail entities work out so well
- OTOH the whole network is quite complex; quite a wonder that backprop through such hostile functions works so well to depth $O(T)$
- Many potential bad choices for $\boldsymbol{A}, \boldsymbol{B}, N$; would be good to know how robust the design is

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] A. Sil and A. Yates, "Re-ranking for joint named-entity recognition and linking," in *CIKM*, 2013, pp. 2369–2374. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.398.9086&rep=rep1&type=pdf

[11] 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

[12] 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

[13] 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/

References (4)

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

[15] 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

[16] 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

[17] X. Cheng and D. Roth, "Relational inference for wikification," in *EMNLP Conference*, 2013, pp. 16–58. [Online]. Available: https://www.aclweb.org/anthology/D/D13/D13-1184.pdf

[18] M. Ponza, P. Ferragina, and S. Chakrabarti, "A two-stage framework for computing entity relatedness in wikipedia," in *CIKM*, 2017, pp. 1867–1876. [Online]. Available: https://dl.acm.org/citation.cfm?id=3132890

References (5)

[19] 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

[20] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to Wikipedia," in *ACL Conference*, ser. ACL/HLT, Portland, Oregon, 2011, pp. 1375–1384. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002642

[21] 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

[22] 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

References (6)

[23] 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

[24] 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

[25] 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

[26] 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 (7)

[27] 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

[28] 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

[29] 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

[30] 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 (8)

[31] 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

[32] 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

[33] 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

[34] 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 (9)

[35] 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

[36] 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

[37] 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

[38] 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 (10)

[39] 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

[40] 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

[41] 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

[42] 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 (11)

[43] 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

[44] 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

[45] 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

[46] 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 (12)

[47] 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

[48] 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

[49] 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

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

References (13)

[51] 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

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

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

[54] 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

[55] 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 (14)

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

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

[58] 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

[59] 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 (15)

[60] 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

[61] 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/

[62] 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

[63] 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/

References (16)

[64] 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

[65] 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

[66] 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

[67] 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

References (17)

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