Cross-Document Coreference Resolution using Latent Features

Axel-Cyrille Ngonga Ngomo[1], Michael Röder[1,2], Ricardo Usbeck[1,2]

[1]University of Leipzig, Germany

[2]R & D, Unister GmbH, Germany

October 20, 2014

Motivation

Google
2 MILLION Searches
Spotify
14 New Songs Added
DOMAINS
70 New Registered
15 THOUSAND Tracks downloaded from iTunes
WordPress
347 New Blog Posts
571 New Created
WEBSITES
Transactions 17 THOUSAND
Walmart
New Tumblr Photos 20 THOUSAND
tumblr.
emails sent 204 MILLION
EMAIL
photos every second 3,600
Instagram
Active Users 11 THOUSAND
Pinterest
Professional searches 11 THOUSAND
Linked in
Amazon Sales $83,000
amazon.com
104 THOUSAND Photos Shared
snapchat
278 THOUSAND Tweets
twitter
20 MILLION Photo Views
flickr
1.4 MILLION Minutes Connecting with Each other
skype
1.8 MILLION likes
350GB of data
41 THOUSAND posts every second
facebook
72 HOURS of Video uploaded
YouTube

Motivation

Problem

- Real-world entities mentioned using very different labels
- Homonyms, e.g., golf

$\Rightarrow$ Simple URI generation for novel entities does not work

Example

P. Diddy, also known as Sean Combs, gave a concert in Harlem today.

Motivation

Problem

- Real-world entities mentioned using very different labels
- Homonyms, e.g., golf

⇒ Simple URI generation for novel entities does not work

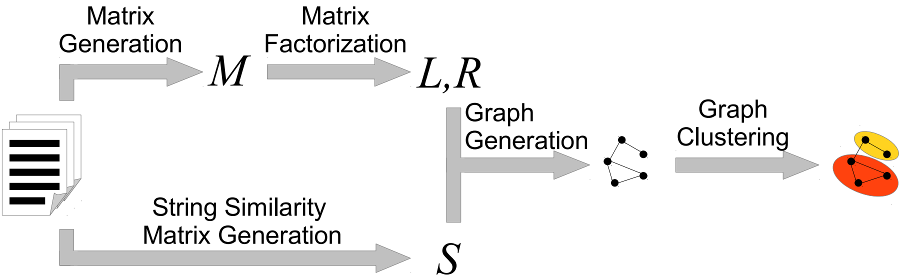
Example

P. Diddy, also known as Sean Combs, gave a concert in Harlem today.

Goal CDCR

Assign the same URI to different mentions of the same real-object even across documents.

Overview



Matrix Generation
M
Matrix Factorization
L,R
Graph Generation
Graph Clustering
String Similarity Matrix Generation
S

Matrix Generation

Idea

Represent entities by multisets that describe the window of words around them.

Matrix Generation

Idea

Represent entities by multisets that describe the window of words around them.

Example

`Yesterday, VW's CEO presented the new Golf in Munich.`

- Stopwords are removed, i.e., {`the`, `in`}
- Window size $\sigma = 1$
- `Golf` is represented by the multiset {`new` (1), `Munich` (1)}
- Within the vector space (`presented`, `new`, `Munich`, `Germany`), this mention has the vector representation $(0, 1, 1, 0)$.

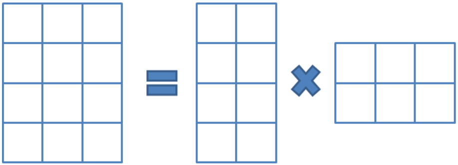
Matrix Generation

word	g_1	g_2	g_3	g_4	g_5
presented	2	1	0	1	0
new	2	0	0	0	1
Munich	2	0	0	0	1
Germany	0	1	1	0	0

Idea

Characterize mentions by using latent features to achieve better comparability

- Output of generation is matrix $M(n, m)$
- Use factorization to compute $L(n, \rho)$ and $R(m, \rho)$ such that $M \approx LR^\top$.
- $\rho \in \mathbb{N} \backslash \{0\}$ is the *rank* of the factorization



- Approach: Minimize $||E||_F^2 = ||M - RL^\top||_F^2 - \frac{\lambda}{2}(||R||_F^2 + ||L||_F^2)$
- Use random initialization and gradient descent approach to update the matrices L and R
- Reduce the error by updating each l_{ik} resp r_{jk} iteratively as follows:

$$l_{jk} \leftarrow l_{jk} - \alpha \frac{\partial e_{ij}}{\partial l_{jk}} = l_{jk} + \alpha \left(2 \sum_{i=1}^{n} e_{ij} r_{ik} - \lambda l_{jk} \right) \quad (1)$$

and

$$r_{ik} \leftarrow r_{ik} - \alpha \frac{\partial e_{ij}}{\partial r_{ik}} = r_{ik} + \alpha \left(2 \sum_{j=1}^{j} e_{ij} l_{jk} - \lambda r_{ik} \right) . \quad (2)$$

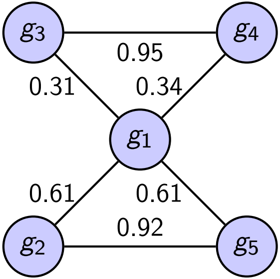
$$M = \begin{pmatrix} 2 & 2 & 2 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}. \quad (3)$$

For $\rho = 2$, our approach computes

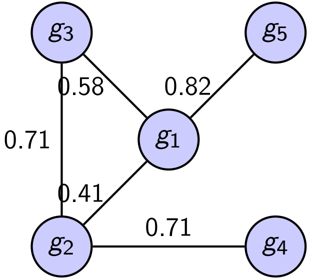
$$L = \begin{pmatrix} 1.385 & 1.102 \\ -0.006 & 0.501 \\ 0.079 & -0.051 \\ -0.234 & 0.712 \\ 0.933 & -0.168 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 0.331 & 1.406 \\ 1.059 & 0.446 \\ 1.118 & 0.363 \\ 0.062 & 0.066 \end{pmatrix}. \quad (4)$$

Graph Generation

- Generate a similarity graph $G = (V, E, w)$
- Set of vertices V is the set of entity mentions
- $w : V \times V \rightarrow [0, 1]$ with $w(v_i, v_j) = s_{ij} \times \frac{l_{(i,\cdot)} \cdot l_{(j,\cdot)}}{||l_{(i,\cdot)}|| \times ||l_{(j,\cdot)}||}$
- Edge is set between v_i and v_j iff $w(v_i, v_j) \geq \theta \in [0, 1]$



(a) Graph generated using $\rho = 2$ and $\theta = 0.3$

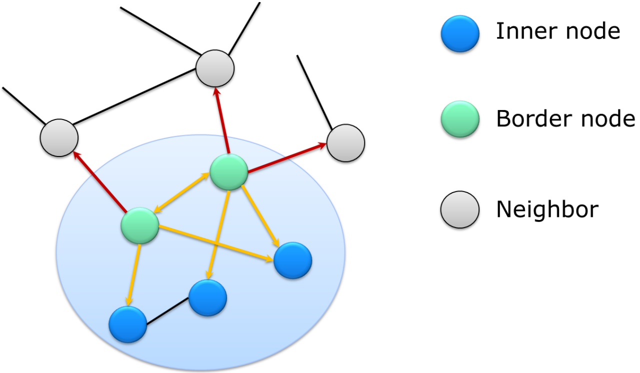


(b) Graph generated using M instead of L and $\theta = 0.3$

Graph Clustering

Borderflow

- Use each vertex as seed.
- Maximize $bf(C) = \frac{\Omega(b(C), C)}{\Omega(b(C), V \setminus C)}$
- Follow iterative approach



Experimental Setup

Goal

Measure effect of latent features on CDCR

- **Baseline**: Use M instead of L
- Measure the influence of
  - the rank ρ
  - the window size σ
  - the hardening
- Settings:
  $\theta = 0.1, \lambda = 0.02, \alpha = 0.0002$

Experimental Setup - Datasets

We use the three corpora of the N^3 collection [1].

	News-100	Reuters-128	RSS-500
Documents	100	128	500
Tokens	48199	33413	31640
Entities	362	444	849
Mentions	1655	880	1000

Results - Influence of rank

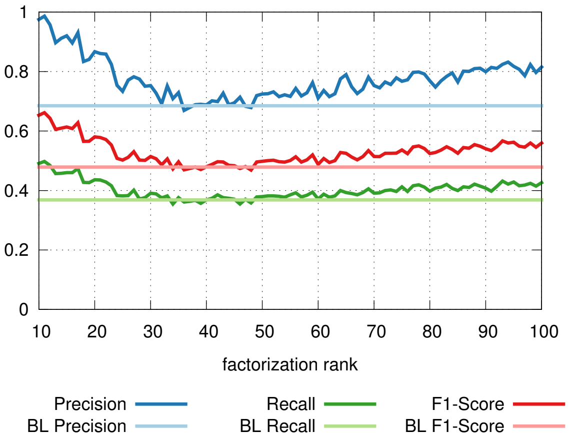


Figure: Precision, recall and F1-score of our approach on the Reuters-128 dataset with different ranks compared to the baseline (BL).

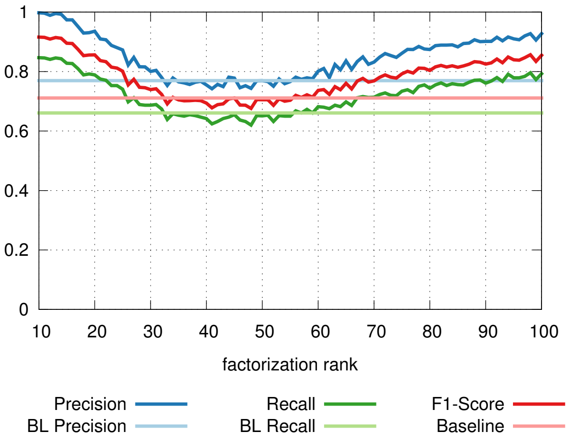


Figure: Precision, recall and F1-score of our approach on the RSS dataset with different ranks compared to the baseline (BL).

Results - Influence of window size

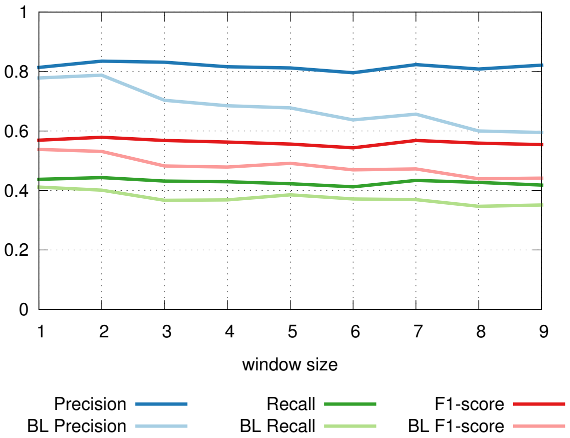


Figure: Precision, recall and F1-score of our approach on the Reuters-128 dataset with different window sizes compared to the baseline (BL).

Results - Influence of window size

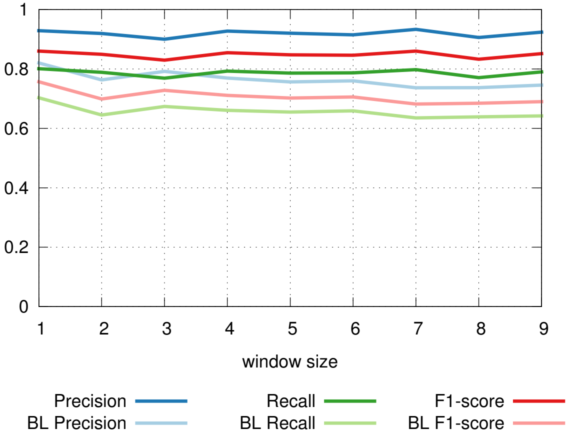


Figure: Precision, recall and F1-score of our approach on the RSS dataset with different window sizes compared to the baseline (BL).

Table: Best improvements in F-measure of our approach (OA) over the baseline (BL)

	Flow Max.		**Set-Based**		**Silhouette**	
	BL	OA	BL	OA	BL	OA
News-100	25.86	**32.21**	23.87	**28.81**	26.56	**34.05**
Reuters-128	47.89	**66.16**	47.00	**56.65**	47.59	**59.60**
RSS-500	71.11	**91.62**	69.57	**85.71**	68.97	**88.22**

Conclusion

- Presented a CDCR approach based on latent features
- Our approach outperforms the baseline by approx. 10%
- Results can be used for better URI generation
- Future work includes
  - Improving convergence of the approach
  - Including knowledge from Linked Data into M

Thank you!
Questions?

Axel Ngonga
University of Leipzig
AKSW Research Group
Augustusplatz 10, Room P616
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de
http://github.com/AKSW/CoreferenceResolution

References

Michael Röder et al. "N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format". In: *The 9th edition of the Language Resources and EvaluationConference, 26-31 May, Reykjavik, Iceland*. 2014. URL: http://svn.aksw.org/papers/2014/LREC_N3NIFNERNED/public.pdf.