A graph-based collective linking approach with Group Co-existence Strength

Chinmay Choudhay, Colm O'Riordan

National University of Ireland (NUI), Galway
{c.choudhary1,colm.oriordan}@university.ie

Abstract. This paper addresses a drawback of many existing graph-based collective entity-linking approaches by introducing the new concept of Group Co-existence Strength (GCS). Doing so, this work proposes an approach to the collective linking of text documents which extends an existing recent approach by taking into account GCS for all possible groups of candidate entities along with standard attributes. Preliminary experimental results indicate that the proposed approach leads to performance gains with selected real world data.

1 Introduction

Named Entity Disambiguation (NED) is the process of linking name-mentions in a document to the accurate real-world entities to which they are referring. These entities can be instances of diverse range of categories such as famous-person, place, institution, country, scientific-discovery etc., collectively forming a Knowledge-base (such as Wikipedia).

Entire *Entity Linking (EL)* process comprises of three major steps namely, recognition of ambiguous name-mentions within a document called *Named Entity Recognition (NER)*, identification of candidate entities for each such ambiguous name-mention and disambiguation of these name-mentions by linking each one with most appropriate entity out of all the candidates, each step being distinct broad research area within itself. The research work described within this paper is focused on the final step, thus assumes all name-mentions within a document being correctly demarcated and a set of candidate entities for each such name-mention being identified beforehand.

2 Related Work

[1], [9], [7], [2] are examples of prominent approaches belonging to *individual-linking* category, which link each name-mention individually based on similarity between context of it within document and description of entity, commonly referred as *Compatibility(CP)* whereas [8], [15],[18], [19], [13], [16], [6], [12], [11], [4], [3] [14] are examples of modern *collective-linking* approaches adopting various supervised and unsupervised linking methods. [5] is a prominent graph-based approach that link all name-mentions within single document simultaneously

by considering semantic relationships between various pairs of entities indicating the chances of both entities being referred in a single real-world document depending upon how closely they are associated with common topic or field, referred as *Semantic Relatedness (SR)* along with CP, which is directly extended within this paper.

3 Drawback of contemporary approaches

Most graph-based collective linking approaches consider SR between all possible pairs of entities that are candidates of two distinct name-mentions appearing within the document for computation of overall linking. Thus for a set of entities associated with entire document consisting of members such that each member is a candidate of single distinct name-mention, value of score indicating the suitability of the entire set being appropriate collective link is computed as a function of SR scores for all possible pairs that can be extracted from the particular set. There is an inherent assumption with this approach which can be stated as follows.

All members of a set of entities have higher chances of being referred together in a single real-world document, if most of the pairs extracted from the set possess strong semantic relationship.

But this assumption does not always hold true, specifically if there is an outlier in the group thus limiting the accuracy of system. Consider text-document stated as Example 1 consisting of four name-mentions namely *Donald*, *Hillary*, *Fox* and *America*.

Example 1. ***Donald*** *will direct the upcoming movie from* ***Fox*** *with* ***Hillary*** *playing lead role in it. The movie will be released across* ***America*** *by the end of 2017.*

Let there be two candidate collective links of entire document namely W_1 and W_2 listed as follows.

W_1 = `[ Donald Trump, Hillary Clinton, Fox, United States of America ]`
W_2 = `[ Donald Petrie, Hillary Swank, Fox Studios, North America ]`

Here by common-sense and real-world knowledge it is evident that W_2 is more appropriate link than W_1 but modern approaches would still link W_1 as most pairs of entities extracted from W_1 have stronger semantic relationship as compared to their counterparts in W_2 (for example pair {*Donald Trump,Hillary Clinton*} as compared to pair {*Donald Petrie and Hillary Swank*} etc.).
To address this issue the paper introduces a new concept called *Group Coexistence Strength (GCS)* as section 4 and proposes an NED approach taking it into consideration as section 5.

4 Group Co-existence Strength

GCS of a group of entities, indicate the chances of all its members being co-referred within any given real-world document. This strength depends on how symmetrically the entities are distributed with respect to each other in terms of mutual SR scores. One way to demonstrate this distribution is to plot all members of entities on a graph with co-ordinates of each being determined by values of Semantic Distance (computed as a factor of SR) of it from pre-decided benchmark members of the same group. Groups with members being more compactly plotted can be considered to be semantically stronger.For sets of candidate collective-links W_1 and W_2 outlined for Example 1 the distribution plots are represented as figure 2 and figure 1 respectively. It is evident from the figures that W_2 is more compactly distributed as compared to W_1 which has an anomaly.

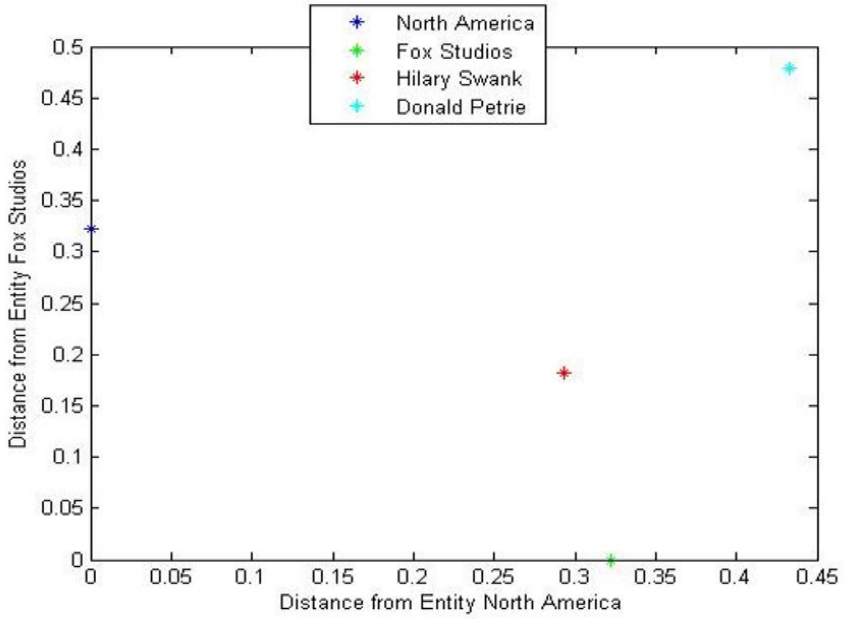

Fig. 1. Distribution of entities in the set W_2

GCS of any group of entities is indicated by value of indicator called *Group Strength Factor (GSF)* described as section 4.1

4.1 Group Strength Factor

GSF for a given set of entities of any size is the minimum value obtained out of all Gaussian values achieved at the positions of all entities within the set,

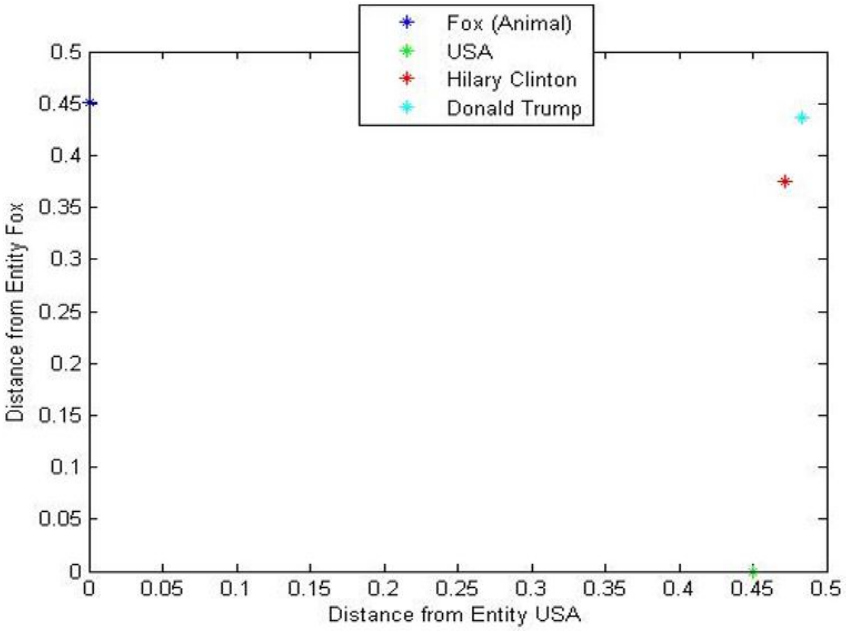

Fig. 2. Distribution of entities in the set W_1

with peak of Gaussian being at average position and standard deviation being a fixed value vector (of size equal to total number of co-ordinates). For a set of entities S let R be set of reference entities such that $R \subset S$. Then for any entity $E_i \in S$, the position of it is defined by equal number of co-ordinates as the size of R, with any j^{th} co-ordinate being computed with respect to j^{th} member of R ($0 \leq$ j $\leq$ Size of R) by equation 1 with SR being Semantic Relatedness [8] and NF being a pre-determined *Normalization Factor*. It is important to note that value of NF would not have any impact on overall performance, provided it satisfies the necessary condition of being common during entire linking-process. It merely exists to provide facility to manage and keep positions of entities as well as GSF values obtained on each, within a considerable range (in possible cases when GSF values would go too low to be easily distinguished, compared or distinctly plotted on graph) to be conveniently analysed.

$$E_{i_{coordinate_j}} = NF * SR(E_i, R_j) \tag{1}$$

For experiments described in this paper NF is considered to be one.
Co-ordinates determining position of peak is given as average of values of same co-ordinate for all Entities belonging to set S computed through equation 2.

$$Peak_{Coordinate_j} = \frac{\sum_{AllE \in S}(E_{coordinate_j})}{n} \tag{2}$$

Having positions of all entities belonging to S and peak (as values of representing co-ordinates), GSF score of S is determined by applying equation 3 with $\mathcal{N}$ representing *Normal distribution* of position of entities belonging to S, around the Peak (computed through equation 2) with a fixed value of standard deviation (σ).

$$GSF = \min_{All E \in S} (E \sim \mathcal{N}(Peak, \sigma)) \tag{3}$$

For the purpose of experimentation, surely larger the size of R would have more accurate positioning (with more co-ordinates), thus more accurate final-linking, though at the cost of lower time-bound efficiency. Once having decided the size of R to be considered during entire linking process, any sub-set of S of that size being used as R would give similar results as computation of GCF involves symmetry of mutual distribution of all entities with respect to each other.
Within this paper the value of σ is randomly considered to be 0.1 (As it does not matter what specific value is taken provided it is same for all examples during both training and testing) whereas size of R is considered to be 1, thus all entities being plotted on 1-D axis.
Basic intuition behind GSF is simply the fact that Gaussian value obtained on the outliers will be relatively much lower as they are positioned at a considerable distance from the peak on the overall plot, thus penalizing the entire group.

5 Linking Approach

The overall collective linking of name-mentions in a given document (let having N name-mentions) simultaneously while taking into account sum of GSF values of all possible groups of entities of a particular size (referred to as GSF_n with n being the size), for all possible sizes greater than two that can be extracted from set of entities being candidate collective-link, involves computation of a term called *Linking Factor (LF)* for all such candidates by applying heuristic formula stated as equation 4. LF value of a particular candidate collective-link (as a set of N entities corresponding to each name-mention) depends upon GSF_n ($2 < n \leq N$) along with sum of Semantic Relatedness scores ($\sum SR$) between all possible pairs of entities as well as sum of values of Compatibility scores ($\sum CP$) between all name-mentions and their respective candidate entities (forming candidate collective link).

$$LF = \sum_{n=3}^{N} (\phi_1 * n^3 + \phi_2 * n^2 + \phi_3 * n + \phi_4) * GSF_n$$

$$+\varphi_2 * \sum SR + \varphi_1 * \sum CP \tag{4}$$

Here N is the total number of name-mentions appearing within document while $\phi_1, \phi_2, \phi_3, \phi_4$, φ_1 and φ_2 are parameter that can be learnt using a set of training dataset. Equation 4 is formulated based on intuition that impact of GCS for groups of all sizes on collective-linking process should not be same,

thus GSF_n being normalized by a quadratic equation of n with optimum degree 3 to avoid both *under-fitting* and *over-fitting*. For a given document consisting of a set of name-mentions appearing within it and a group of sets of entities as candidate collective-links (having equal number of entities as name-mentions with each entity being associated with single distinct name-mention), LF score of all such candidates can be computed to identify the one with maximum score as most appropriate.

6 Experimentation

As already explained in section 1, the proposed approach can be applied to rank respective candidates of all name-mentions appearing in a single document for the purpose of collectively linking all such name-mentions to their respective most appropriate entities simultaneously. Thus dataset utilized for the purpose of training and testing the approach should consist of text-documents with all name mentions demarcated and candidates for each being identified beforehand. Section 6.1 describes the structure and process of generation of final datasets whereas subsequent sections elaborate on computation, training and testing procedures.

6.1 Dataset

First phase of experimentation involves extraction of information from original IITB helpfulness dataset [17] to formulate three distinct final datasets to be utilized for final training and testing of proposed approach. IITB dataset is comprised of a collection of text-documents related to varied range of subjects such as sports, science, politics etc. with details of all name-mentions within all documents including Title of correct Wikipedia link to be each, is represented a single large JSON document. As proposed approach identifies most appropriate collective link of all name-mentions simultaneously after learning the parameters of equation 4, three distinct final datasets having most suitable specific structure are created by modifying original dataset through elaborate process. Following two sub-sections describe *Structure* and *Process of creation* of Final Datasets respectively.

Final datasets As already explained final training requires parameters of equation 4 to be learnt through Logistic Regression which fundamentally requires a set of positive and negative training example for its implementation. Final datasets are constituted by such examples with each having label as either positive or negative with a single example consisting of a collection of top 100 name-mentions appearing in a single text-document ranked according to their relevance, with each being paired up with one of its candidate entities. Examples with all name-mentions being paired to their respective correct links as per information provided with original IITB helpfulness datasets JSON document can be considered as having positive label while others as having negative labels.

Various text-documents within IITB dataset contain varied number of name-mentions being appeared in the content, with minimum number being as 100. Thus for all the documents only top 100 high-relevance name-mentions are being considered while ignoring others, for the purpose of maintaining homogeneity between all examples of datasets, essential for training and testing convenience. Relevance of each name-mention within specific text-document for the purpose of collective linking is indicated within original IITB dataset as *relevance index*. For each text-document all name-mentions appearing within it are sorted according to*relevance index* and top 100 members are retained while ignoring others.
Characteristic feature that mainly distinguishes three datasets is the degree of overlap among examples contained by each one of them. For a collection of sets of entity-mention pairs forming a single dataset overlap of that dataset refers to the percentage of common members belonging to any two given candidate sets of entities that can possibly be a collective link of single common text-document. Details of all three datasets is summarized as Table 1.

	Total Number of examples	Number of Positive examples	Number of Positive examples	Percentage of Overlap
Dataset 1	9161	97	9064	More than 90%
Dataset 2	194	97	97	Approximately 30%
Dataset 3	194	98	96	Less than 10%

Table 1. Information about Datasets

Process of creation of Datasets As the proposed approach identifies most appropriate candidate entity to be linked to each of the name-mentions within single document simultaneously, evaluation of it requires at least one incorrect and one correct candidate entity that can be linked to each name-mention within all text-documents. All name-mentions are provided by the correct link as Wikipedia title within original IITB dataset while incorrect candidate is extracted from *See Also* section of that correct link. All other Wikipedia page hyper-links within *See Also* section of Wikipedia article of a correct link of given name-mention are sorted according of similarity of Bag of Words (BOW) extracted from these with BOW extracted from contents of correct link in decreasing order. Hyper-link of Wikipedia page on the top of the list is considered as second incorrect candidate of particular name-mention.
Having two candidates for each name-mentions, final datasets are created by pairing up these name-mentions with each of its candidates and re-arranging all pairs with name-mentions appearing in single text-document as a large collection. Each such collection being a single possible collective link of the particular text-document forms single example with label as positive if all name-mentions are paired with correct link and negative otherwise. Three distinct methods adopted to perform re-arrangement classifies three distinct datasets.

6.2 Computation

For a given set of entity-mention pair being a possible collective link, Equation 4 computes Linking factor by taking into account three distinct parameters namely sum of *Semantic Relatedness (SR)* scores of all possible pairs of entities that can be extracted from the set, sum of *Compatibility (CP)* scores of all possible entity-mentions pairs forming the set and sums of *Group Strength Factors (GSF)* of all possible group of entities of a specific size n (`n>2`) that can be extracted from set, for all possible values of n. The processes adopted for computation of these scores are explained as follows.

1. Compatibility (CP) :
It is computed between context of name-mention and entity description. For this experimentation context of a name-mention is considered as twenty words before and after it within text-file content and entity description is simply the content of respective Wikipedia article. For a name-mention NM and a Wikipedia entity W, let B_{NM} and B_W be Bags of N-grams extracted from their context and description respectively with value of N ranging from 1 to 3.
Compatibility between NM and W is given by equation 5.

$$CP(NM, W) = TFIDF_{NM} * {V_{W/NM}}^T \tag{5}$$

Where $TFIDF_{NM}$ consists of TFIDF scores of all N-grams within B_{NM} with respect to all the text-documents within original IITB dataset. $V_{W/NM}$ is a Boolean vector of length equal to length of B_{NM} with values obtained from equation 6.
For all i= 1 to length of $V_{W/NM}$

$$V_{W/NM_i} = \begin{cases} 1 \text{ if } B_{NM_i} \in B_W \\ 0 \quad \text{otherwise} \end{cases} \tag{6}$$

2. Semantic Relatedness (SR) :
There are numerous approaches to compute Semantic Relatedness between Wikipedia Entities but the most common one is proposed within [10] which uses the intersection and union of hyper-links shared between two given entities for computation by applying equation 7. For this experimentation same-method is utilized adopting common practice.

$$SR(x, y) = 1 - \frac{\log(\max(|X|, |Y|)) - \log(X \cap Y)}{\log |W| - \log(\min(|X|, |Y|))} \tag{7}$$

The components of formula are described as follows.

- $|W|$: Total Number of articles on Wikipedia
- $|X|$: Number of hyperlinks on entity x
- $|Y|$: Number of hyperlinks on entity y

– $X \cap Y$: Number of hyperlinks shared by entities x and y

3. Group Strength Factor (GSF):
For any group of entities of size greater than two GSF is computed by applying Equation 3. Ideally application of equation 4 for computation of Linking factor (LF) for a given set of entity-mention pairs requires GSF values of all possible groups of entities of size three or more that can be extracted from the set being taken into account. Since there are 100 entities within each example, total number of GSF computations that need to be performed for each example is as follows.

$$\binom{100}{100} + \binom{100}{99} + \binom{100}{98} + \binom{100}{97} + ... + \binom{100}{3}$$

This reduces the time-efficiency of overall training and testing to extremely low, thus making the evaluation of hypothesis in stipulated time-period infeasible. Considering this limitation, for the purpose of this experimentation GSF for the groups of entities with maximum size as 10 only is taken into consideration. Maximum size is considered to be 10 because it is the maximum value for which experimentation process held feasibility within decided time-constraint.

6.3 Final Matrix

As explained in section 6.2 a single example of final dataset is formed by collections of all name-mentions (top 100 based on relevance in case of this particular experimentation) appearing in a single text-document, with each being paired-up with one of its candidate entities. For each such example all 10 distinct values namely sum of SR values, sum of CP values and sums of GSF values of group of size ranging from 3 to 10 are represented as single 1*10 vector. Thus an example e is represented as vector V_e given be equation 8.

$$V_e = [\ GSF_{10}\ GSF_9\ ...\ GSF_3\ SR\ CP\ ] \tag{8}$$

Thus entire dataset consisting of m examples can be represented as an m*10 matrix M_d given by equation 9 and an m*1 Boolean vector holding labels of all m examples.

$$M_d = [\ V_{e_1}\ V_{e_2}\ .....\ V_{e_m}\ ]^T \tag{9}$$

6.4 Training and Testing

Final Collective linking is performed by computing *Linking Factor (LF)* for each training example given by formula described as equation 2.2. For the case of current experimentation process, since maximum size of group of entities is considered to be 10, the formula can written as equation 10.

$$LF = (10^3 * \alpha_1 + 10^2 * \alpha_2 + 10 * \alpha_3 + \alpha_4) * \sum_{i=1}^{m} SSF_{10}$$

$$+(9^3 * \alpha_1 + 9^2 * \alpha_2 + 9 * \alpha_3 + \alpha_4) * \sum_{i=1}^{m} SSF_9 + ...+$$

$$(3^3 * \alpha_1 + 3^2 * \alpha_2 + 3 * \alpha_3 + \alpha_4) * \sum_{i=1}^{m} SSF_3 +$$

$$\theta_1 * \sum_{i=1}^{m} SR + \theta_2 * \sum_{i=1}^{m} CP \tag{10}$$

Value of *Linking factor (LF)* for all examples within a given dataset d can be represented as a single m*1 matrix called *LFMatrix*. After performing mathematical derivations on equation 10 it can be proved that *LFMatrix* of d can be computed by applying equation 11.

$$LFMatrix = (M_d * Multiplier) * P \tag{11}$$

Here M_d is matrix defined as equation 8. P and Multiplier as given as equations 12 and 13. P is the parameter matrix that needs to be learnt through Logistic Regression. It is initialized with random values and is subsequently updated after each iteration until optimization.

$$P = [\ \alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \theta_1\ \theta_2\ ]^T \tag{12}$$

$$Multiplier = \begin{bmatrix} 10^3 & 10^2 & 10 & 1 & 0 & 0 \\ 9^3 & 9^2 & 9 & 1 & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 3^3 & 3^2 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

To evaluate performance of proposed approach, given dataset is split in the ratio of 60% and 40%, with first 60% examples being utilized to learn parameters within equation 4 (represented as single matrix P in equation 12) through Logistic Regression whereas the testing is performed on last 40% of dataset. Training and testing is performed on all three datasets distinctively and results obtained by each are discussed as section 7.

7 Preliminary Results and Future Work

Having probability matrix for a given test-dataset as described in section 6.2, considering a fixed threshold value of 0.5, predictions are made for each example thus obtaining a predicted Boolean matrix to be compared with actual Boolean matrix. Table 2 compares the average results achieved on three

datasets with results of Wikification approach [9] and approach [5] which are benchmark individual-linking and collective-linking graph-based approaches respectively. Though the results are yet to be compared with various state of the art approaches, preliminary results indicate that proposed approach performed significantly better than both benchmark approaches.

	Average Precision	**Average Recall**	**Average F-Score**
Wikify	0.55	0.32	0.38
Collective Graph-based	0.69	0.76	0.73
Our-approach	0.69	0.996	0.81

Table 2. Comparison of results between various approaches

Future work would include much more exhaustive testing and evaluation of proposed approach on larger datasets.

References

1. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Eacl. vol. 6, pp. 9–16 (2006)
2. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285. Association for Computational Linguistics (2010)
3. Ganea, O.E., Ganea, M., Lucchi, A., Eickhoff, C., Hofmann, T.: Probabilistic bag-of-hyperlinks model for entity linking. In: Proceedings of the 25th International Conference on World Wide Web. pp. 927–938. International World Wide Web Conferences Steering Committee (2016)
4. Hachey, B., Radford, W., Curran, J.R.: Graph-based named entity linking with wikipedia. In: WISE. pp. 213–226. Springer (2011)
5. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 765–774. ACM (2011)
6. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 782–792. Association for Computational Linguistics (2011)
7. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging wikipedia semantics. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 179–186. ACM (2008)
8. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 457–466. ACM (2009)

9. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242. ACM (2007)
10. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 509–518. ACM (2008)
11. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics 2, 231–244 (2014)
12. Naderi, A.M.: Unsupervised entity linking using graph-based semantic similarity (2016)
13. Pan, X., Cassidy, T., Hermjakob, U., Ji, H., Knight, K.: Unsupervised entity linking with abstract meaning representation. In: HLT-NAACL. pp. 1130–1139 (2015)
14. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 365–374. ACM (2017)
15. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 1375–1384. Association for Computational Linguistics (2011)
16. Singh, S., Riedel, S., Martin, B., Zheng, J., McCallum, A.: Joint inference of entities, relations, and coreference. In: Proceedings of the 2013 workshop on Automated knowledge base construction. pp. 1–6. ACM (2013)
17. Yamada, I., Ito, T., Usami, S., Takagi, S., Takeda, H., Takefuji, Y.: Evaluating the helpfulness of linked entities to readers. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media. pp. 169–178 (2014)
18. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 483–491. Association for Computational Linguistics (2010)
19. Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., Gaffney, S.: Resolving surface forms to wikipedia topics. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1335–1343. Association for Computational Linguistics (2010)