Combining Heuristics and Learning for Entity Linking

Hien T. Nguyen(✉)

Ton Duc Thang University, Ho Chi Minh, Vietnam
hien@tdt.edu.vn

Abstract. Entity linking refers to the task of mapping name strings in a text to their corresponding entities in a given knowledge base. It is an essential component in natural language processing applications and a challenging task. This paper proposes a method that combines heuristics and learning for entity linking by (i) learning coherence among co-occurrence entities within the text based on Wikipedia's link structure and (ii) exploiting some heuristics based on the contexts and coreference relations among name strings. The experiment results on TAC-KBP2011 dataset show that our method achieves performance comparable to the state-of-the-art methods. The results also show that the proposed model is simple because of using a classifier trained on just two popular features in combination with some heuristics, but effective.

Keywords: Entity linking · Entity disambiguation · Wikification

1 Introduction

The task of identifying *surface forms* – strings used mentioning to entities in text – and linking them to their corresponding knowledge base (KB) entries that provide background information about the referent entities is an essential component in natural language processing applications. This task was well-known as entity disambiguation [1] or entity linking [2]. When the used KB is Wikipedia, the task is also known as wikification [5]. This work pursues the entity linking task that is to annotate/map surface forms in text with/to their corresponding entries in a given KB, i.e., Wikipedia. For instance, given the mention *Jim Clark* in the context "Jim Clark took pole position for the Monaco Grand Prix", a good entity linking method will recognize Jim Clark as the British Formula One racing driver, but in another context "Netscape cofounder Jim Clark returns to the Forbes Billionaires List", that method will recognize Jim Clark as the cofounder of Netscape. From now on, we use entity linking in place of both entity disambiguation and wikification.

Entity linking (EL) is challenging due to surface forms ambiguity. That is because one surface form may refer to different entities in different occurrences and one entity may be referred to by different surface forms in different contexts. For example, the surface form *Michael Jordan* in different occurrences may refer to the basketball player (who had ever played for Chicago Bulls), the professor working at UC Berkeley, etc.; or surface forms *Michael Jordan* and *Jordan* in different contexts can referred to the same person. In particular, given a document d, let $S = \{s_1, s_2, \ldots, s_N\}$ be

P.C. Vinh et al. (Eds.): ICCASA 2013, LNICST 128, pp. 369–378, 2014.
DOI: 10.1007/978-3-319-05939-6_36,

the set of surface forms in d; the goal is to produce annotations of the set of surface forms with the set of KB entries $A = \{a_1, a_2, \ldots, a_N\}$. When the used KB is Wikipedia, A is a set of Wikipedia articles.

Since 2009, the entity linking shared task yearly held at Text Analysis Conference (TAC) [2] has attracted more and more attention of research groups all over the world and many approaches to entity linking have been proposed. In TAC entity linking, given a query consisting of a surface form and a background document where the surface form occurs, the EL system is required to provide the identifier (ID) of the KB entry of that surface form; or *NIL* if there is no such KB entry [2]. Figure 1 shows an example in which *Georgia* is the surface form targeted disambiguation. The figure also shows that co-occurrence name strings such as "US" and "Atlanta" actually help to clarify which entity Georgia actually refers to.

In this paper, we propose an entity linking method that tries to model how human beings disambiguate a surface form. When reading a text and encountering a surface form, one may rely on his/her knowledge accumulated in the past and the context of the text to identify which one is the underlying entity of that surface form. Indeed, our method exploits prior knowledge about entities and analyzes the context to perform linking decisions. Our proposed model was presented in Fig. 2 with three key steps: (1) candidate generation, (2) linking by heuristics, and (3) linking by learning. As showed in Fig. 2, our entity linking system receives an input as a query that consists of a surface form and the document where that surface form occurs, and then outputs the ID of the KB entity that the surface form actually mentions.

The contribution of this paper is three-fold as follows: (i) we propose a model that combines heuristics and learning for entity linking; (ii) we show that the proposed model is simple with several heuristics and a classifier trained on just two popular features, but effective in that it gives performance comparable to the state-of-the-art methods; and (iii) we evaluate the proposed method on a public dataset and show that it gets good performance.

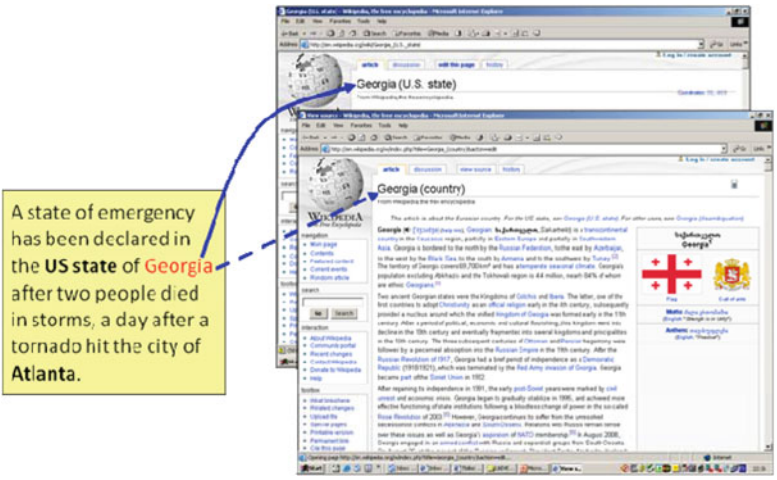

Fig. 1. Wikification

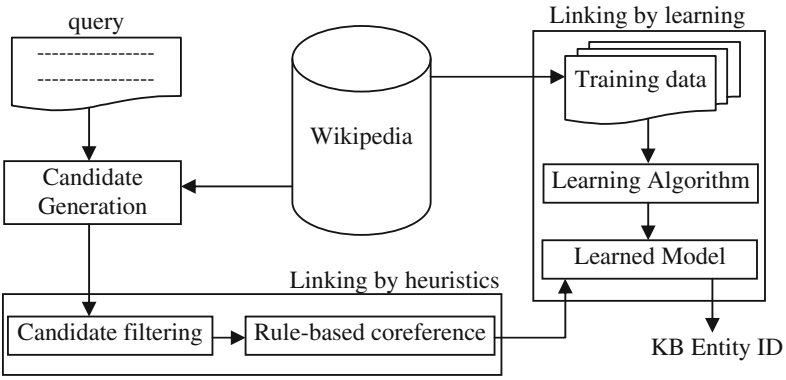

Fig. 2. Our entity linking system

The rest of this paper is organized as follows. Section 2 presents how our method generates candidates for a surface form. Sections 3 and 4 present linking by heuristics and learning respectively. Experiments and results are presented in Sect. 5. Section 6 presents related work. Finally, we draw conclusion and perspectives for future work.

2 Candidate Generation

Wikipedia is a free encyclopedia written by a collaborative effort of a large number of volunteer contributors. There are many meaningful resources of information in Wikipedia that we can exploit for entity linking. Our method proposed in this paper exploits the following resources.

Articles. A basic entry in Wikipedia is an *article* that defines and describes a single entity. It is uniquely identified by its title that is considered as its ID and includes a surface form of that entity. When the surface form is ambiguous, the title may contain further information that we call *title-hint* to distinguish the described entity from others. The title-hint is separated from the surface form by parentheses, e.g. "`John McCarthy (computer scientist)`", or a comma, e.g. "`Columbia, South Carolina`".

Categories. The category hierarchy of Wikipedia is a kind of collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Each article in Wikipedia belongs to some categories. For instance, from the categories of the article describing `John McCarthy (computer scientist)` in Wikipedia, we extract its category labels as follows: Stanford University faculty; Lisp programming language; Artificial intelligence researchers; etc.

Redirect Pages. A redirect page typically contains only a reference to an article. The title of a redirect page is an alternative surface form of the described entity or concept in that article. For example, from redirect pages of the United States, we extract alternative surface forms of the United States such as "US", "USA", "United States of America", etc.

Links. Each page consists of many outgoing links (outlinks) and ingoing links (inlinks). Each link is associated with an anchor text that represents the surface form of the corresponding entity. We collect candidates of a surface form based on outlinks in all articles in which the surface form occurs as labels of the outlinks.

We extract surface forms from the titles of articles and the titles of redirect pages to build a dictionary in which each entry is a surface form. Each entry surface form in the dictionary is mapped to the set of entities that the surface form may denote in Wikipedia. The set of entities is identified by exploiting outlinks in all articles in Wikipedia. As a result, an entity e is included in that set if and only if the surface form can be used to refer to e. Given a surface form, its candidates are retrieved by looking up the dictionary.

3 Linking by Heuristics

Co-occurring entities in a text may have relation with each other. Furthermore, the referent of a surface form can be inferred from nearby entities that have already been identified in the text. For example, when "Michael Jordan" occurs with "Chicago Bulls" or "NBA" in a text, it is more likely that the surface form "Michael Jordan" refers to the former player of Chicago Bulls basketball team. In reality, an entity may have several different surface forms. Therefore, when referring to a certain entity, one can use one or more of its surface forms. We observed that some surface forms co-occurring in a text and referring to the same entity is common. So this method exploits coreference relations among co-occurring surface forms for entity linking.

- Candidate filtering: We employ heuristics, H_1, H_2, and H_3 proposed by Nguyen and Cao (2012) [23] to filter candidates of surface forms. If the number of a surface form is reduced to 1, it is considered as disambiguated and linked to its corresponding entity in the KB.
- Coreference: We employ some of orthomatcher rules proposed in [16] to identify if two certain surface forms are coreferent and then coreference chains among surface forms in the query document are established.

Note that to produce a reliable coreference relation between two mentions, we prohibit the transitive property. That is because in many cases transitivity in coreference relations causes failure. In particular, assume we know that $\{m_1, m_2\}$ and $\{m_2, m_3\}$ are coreferent pairs, we do not imply that m_1 and m_3 are coreferent. An example in [16] showed that assuming transitivity and two coreferent pairs {BBC News, News} and {News, ITV News} imply wrongly that {BBC News, ITV News} are coreferent.

4 Linking by Learning

Milne and Witten [10] employed some classification algorithms to train classifiers using three features namely: commonness (CM), semantic relatedness (SR), and context quality (CQ). The authors showed that Bagged C4.5 give the best

performance. In this section, we re-present features proposed by Milne and Witten [10] for representing entities and how to rank candidate entities of a surface form.

4.1 Commonness

Let s be a surface form, CE_s be a set of candidate entities of s. Commonness [11] of an entity $e \in CE_s$ is the probability of s to link to e. The commonness of a certain pair of the entity e and the surface form s is computed based on Wikipedia as a KB, with different occurrences of s linked to (i.e., referring to) different entities (i.e., Wikipedia articles) including the entity e and defined as follows:

$$Commonnes(e) = \frac{count_s(e)}{\sum_{e_i \in CE_s} count_s(e_i)}$$

where $count_s(e)$ is a function that returns the number of times the surface form s is used to refer to entity e in a certain KB. For instance, assuming that in a KB, a surface form s occurs 10 times and refers to three different entities a, b, c, in which 7 times s refers to a, 2 times s refers to b respectively; then commonness (a) = 7/10 = 0.7, commonness (b) = 2/10 = 0.2, commonness (c) = 1/10 = 0.1; therefore, a is considered as more popular than b and c in the given KB.

4.2 Semantic Relatedness

Given two entities e_1 and e_2, let A_1 be the set of all Wikipedia articles, each of which has a link to e_1, A_2 be the set of all Wikipedia articles, each of which has a link to e_2, and W is the set of all articles in Wikipedia; semantic relatedness between the two entities, e_1 and e_2, called $sem(e_1,e_2)$ is defined as follows:

$$Sem(e_1, e_2) = 1 - \frac{\log(\max(|A_1|, |A_2|)) - \log(A_1 \cap A_2)}{\log(|W|) - \log(\min(|A_1|, |A_2|))}$$

Let E be the set of entities that have already been identified, which are called context entities. We calculate semantic relatedness of an entity e, denoted $SR_w(e)$ and $SR(e)$, as respectively *weighted* average and average of its semantic relatedness to context entities.

$$SR_w(e) = \frac{\sum_{e' \in E} weight(e') x\, Sem(e, e')}{\sum_{e' \in E} weight(e')}$$

$$SR(e) = \frac{\sum_{e' \in E} Sem(e, e')}{|E|}$$

The weight of each entity $e' \in E$ of the surface form s in $SR_w(e)$ was used as the third features in [10] to balance between commonness and semantic relatedness; and is calculated as follows:

$$weight(e) = \frac{\alpha SR(e) + \beta IM(s)}{\alpha + \beta}$$

The $IM(s)$ function estimates the important of the surface form s. We observed that not all surface forms, as well as their referents, play an equally importance role in disambiguation decision of a certain surface form. In other word, some surface forms are more informative than other ones. $IM(s)$ is calculated as follows:

$$IM(s) = \frac{|A'(s)|}{|A(s)|}$$

where $A(s)$ is the set of Wikipedia articles in which s occurs and $A'(s)$ is the set of Wikipedia articles in which s occurs as a label of an outgoing link.

4.3 Linking by Expanding Candidate Set

Our proposed method utilizes coreference relations among surface forms to expand the set of candidates of a certain surface form and then ranks those candidates to choose the best one. It firstly ranks candidates of the surface form to be linked and choose the candidate that having the rank higher a threshold; otherwise, the set of candidates of that surface form is expanded by all candidates of all its coreferent surface forms. For instance, assume that two surface forms s and s' are coreferent and s is the surface form to be linked; let $\{c_1, c_2, c_3\}$ be the set of candidates of s and $\{c'_1, c'_2\}$ be the set of candidates of s'; assume that after being ranked, c_1 has the highest rank and the rank of c_1 is lower than a threshold, our proposed method will rank c'_1 and c'_2 and if the highest rank between those of them is greater than a threshold, s is linked to the corresponding candidate, otherwise, s is linked *NIL*. Note that the detected list of candidates for each surface form might not be complete; therefore, our method does not require two certain referent candidates c_i and c'_j of s and s' respectively must be the same.

Note that the method presented in [10] considers a surface form that has only one candidate as an unambiguous one and the mapping between that surface form and its sole candidate as the final linking decision. However, in reality, the sole candidate of a surface form may not be the entity that the surface form actually refers to. Our proposed method can overcome this drawback by exploiting the coreference relations. Indeed, for a surface form that having only one candidate, our method may link the surface form to an entity other than its sole candidate. For example, assume that two surface forms s and s' are coreferent and s is the surface form to be linked; let c_1 be the sole candidate of s and $\{c'_1, c'_2\}$ be a set of candidates of s'; assume that c'_1 has the highest rank among those of c_1, c'_1, c'_2, our method will link s to c'_1 instead of c_1.

5 Evaluation

We employ the Bagged C4.5 classification algorithm to train a classifier using the features presented in Sect. 4. As in [10], we train our system on a collection of 500 Wikipedia articles and use 100 other Wikipedia articles that do not appear in the

training set to tune parameters. We evaluate our proposed method on TAC-KBP2011 dataset. This dataset consists of 2250 entity mention queries, in which 1124 entity mentions refer to entities described by Wikipedia articles. The evaluation metrics we use are micro-average accuracy (MAA) and B-Cubed+ [2]. We conducted two experiments: (1) without expanding candidate set, namely *Exp1* and (2) with expanding candidate set, namely *Exp2*. Tables 1 and 3 show MAA overall results of *Exp1* and *Exp2* on TAC-KBP2011 dataset respectively. Tables 2 and 4 show B-Cubed+ F1 overall results of *Exp1* and *Exp2* on TAC-KBP2011 dataset respectively. Table 5 show the performance of our method among top 5 best systems in TAC 2011 [2].

The results show that our proposed method is simple, but its performance is comparable to sophisticated methods proposed in top 5 best systems submitted to TAC 2011. Tables 3 and 4 show that expanding candidate set using coreference relations among co-occurrence surface forms improves about 9 % in the best cases when combining commonness and semantic relatedness for training the classifier.

Table 1. The MAA overall results of *Exp1* on TAC-KBP2011 dataset

Feature sets	All (2,250) (%)	NIL (1126) (%)	Non-NIL (1124) (%)
CM	68.3	90.6	46.0
CM+SR	72.7	96.6	48.7
CM+SR+CQ	**73.4**	**94.8**	**52.0**

Table 2. The B-Cubed+ F1 overall results of *Exp1* on TAC-KBP2011 dataset

Feature sets	All (2,250) (%)	NIL (1126) (%)	Non-NIL (1124) (%)
CM	65.5	87.6	44.9
CM+SR	69.6	93	47.3
CM+SR+CQ	**70.4**	**91.4**	**50.6**

Table 3. The MAA overall results of *Exp2* on TAC-KBP2011 dataset

Feature sets	All (2,250) (%)	NIL (1126) (%)	Non-NIL (1124) (%)
CM	75.3	87.8	62.8
CM+SR	**82.5**	**95**	**69.9**
CM+SR+CQ	81.7	92.5	71.0

Table 4. The B-Cubed+ F1 overall results of *Exp2* on TAC-KBP2011 dataset

Feature sets	All (2,250) (%)	NIL (1126) (%)	Non-NIL (1124) (%)
CM	72.7	85	61.5
CM+SR	**79.5**	**91.3**	**68.4**
CM+SR+CQ	78.8	88.9	69.4

Table 5. Our method among top 5 best systems in TAC 2011 [2]

Systems	MAA (%)	B-Cubed+ F1 (%)
LCC [19]	86.1	84.6
NUSchime [18]	86.3	83.1
Ours	**82.5**	**79.5**
Stanford_UBC [20]	79.0	76.3
CUNY [22]	77.8	77.1
CMCRC [21]	77.9	75.4

6 Related Work

To date, many approaches have been proposed for EL using Wikipedia. All of them can fit into three disambiguating strategies: *local*, *global*, and *collective*. Local methods disambiguate each mention independently based on local context compatibility between the mention and its candidate entities using some context features. Global and collective methods assume that linking decisions are interdependence and there is coherence between co-occurrence entities in a text, enabling the use of measures of semantic relatedness for disambiguation. While collective methods simultaneously perform disambiguation decisions, global methods disambiguate mentions in turn.

As a local approach, the method proposed in [12] uses an SVM kernel to compare the lexical context around a certain mention to that of its candidates, in combination with estimating correlation of the contextual words with the candidates' categories. Each candidate of a mention is a Wikipedia article and its lexical context is the content of the article. In [15] the authors implemented and evaluated two different disambiguation algorithms. The first one was based on the measure of contextual overlapping between the local context of a mention and the content of candidate Wikipedia articles to identify the most likely candidate. The second one trains a Naïve Bayes classifier for each mention using three words to the left and the right of outlinks in Wikipedia articles, with their parts-of-speech, as contextual features. In [7] the authors employed classification algorithms that learn context compatibility for disambiguation. The authors in [8] and [9] employed learning-to-rank techniques to rank all candidates and link the mention to the most likely one. The method presented in [4] improved the one proposed in [7] by a learning model for automatically generating a very-large training set and training a statistical classifier to detect name variants. The main drawback of the local approaches is that they do not take into account the interdependence between linking decisions.

Global approaches assumed interdependence between linking decisions and exploited two main kinds of information that are disambiguation context and semantic relatedness. Cucerzan [13] was the first to model interdependence among disambiguation decisions. In [13], a disambiguation context consists of all Wikipedia contexts that occur in the text and semantic relatedness is based on overlapping in categories of candidates, where each candidate corresponds to a mention. Wikipedia contexts are phrases that comprise inlink labels, outlink labels, and title-hints of all Wikipedia articles. The limitation of this approach is to add irrelevant cues to the disambiguation context.

The proposed method in [14] extended the work in [10] by resolving jointly optimization problem of overall disambiguation decisions using two approximation solutions. Ratinov [5] proposed an approach that combines both local and global methods. Kataria [6] proposed a weakly semi-supervised Latent Dirichlet Allocation model for modeling correlations among words and among topics for disambiguation. Sen [1] adapted topic models for EL. His method exploited proximity to learning word-entity association with observations that a word appears closer to a mention to be stronger indicator of its referent. In [17] the authors mined word-entity association for named entity disambiguation.

Han and Sun [3] proposed a collective approach that firstly builds a referent graph for a text based on local context compatibility and coherence among entities and then disambiguates mentions by a collective inference method using the referent graph. A referent graph is a weighted and undirected graph $G = (E, V)$ where V contains all mentions in the text and all possible candidates of these mentions. Each node represents a mention or an entity. The graph has two kinds of edges: (i) A mention-entity edge is established between a mention and an entity and its weight is calculated using cosine similarity implemented in a bag-of-words model as in [12]; and (ii) An entity-entity edge is established between two entities and its weight is calculated using semantic relatedness between these entities. The author adopted the formula presented in [10] to calculate the semantic relatedness between two entities. The collective algorithm collects initial evidence for each mention and then reinforces the evidence by propagating them via edges of the referent graph.

7 Conclusion

Entity linking is an essential task in natural language processing applications such as semantic web, information retrieval, question answering, or knowledge base population. This paper proposes a method that links surface forms in a text to entries of a given knowledge base. The method combines heuristics and learning for entity linking. The method exploits some heuristics based on the contexts and coreference relations, and learns coherence among co-occurrence entities within the text based on Wikipedia's link structure. The experiment results show that our proposed method is simple and effective. The results also show that expanding candidate set using coreference relations among co-occurrence surface forms significantly improves the performance of entity linking systems.

References

1. Sen, P.: Collective context-aware topic models for entity disambiguation. In: WWW 2012 (2012)
2. Ji, H., Grishman, R., Dang, H.T.: An overview of the TAC2011 knowledge base population track. In: Proceedings of Text Analysis Conference (TAC 2011) (2011)
3. Han, X., Sun, L., and Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of SIGIR 2011, pp. 765–774 (2011)

4. Zhang, W., Sim, Y.C., Su, J., Tan, C.-L.: Entity linking with effective acronym expansion, instance selection and topic modeling. In: Proceedings of the 20th IJCAI (IJCAI 2011), pp. 1909–1904 (2011)
5. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to Wikipedia. In: Proceedings of ACL-HLT 2011 (2011)
6. Kataria, S., Kumar, K., Rastogi, R., Sen, P., Sengamedu, S.: Entity disambiguation with hierarchical topic models. In: KDD 2011
7. Zhang, W., Su, J., Tan, C.-L., Wang, W.: Entity linking leveraging automatically generated annotation. In: Proceedings of COLING 2010 (2010)
8. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Proceedings of HLT: NAACL 2010 (2010)
9. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of COLING 2010 (2010)
10. Milne, D. and Witten, I.H.: Learning to link with Wikipedia. In: Proceedings of the 17th ACM CIKM (CIKM 2008), pp. 509–518 (2008)
11. Medelyan, O., Witten, I.H., Milne, D.: Topic indexing with Wikipedia. In: Proceedings of Wikipedia and AI Workshop at the AAAI-2008 Conference (2008)
12. Bunescu, R., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the EACL (EACL 2006), pp. 9–16 (2006)
13. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of EMNLP-CoNLL Joint Conference (EMNLP-CoNLL 2007), pp. 708–716 (2007)
14. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: collective annotation of Wikipedia entities in web text. In: KDD 2009 (2009)
15. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the 16th ACM CIKM, pp. 233–242 (2007)
16. Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., Cunningham, H.: Shallow methods for named entity coreference resolution. In: Proceedings of TALN 2002 Workhop (2002)
17. Li, Y., Wang, C., Han, F., Han, J., Roth, D., Yan, X.: Mining evidences for named entity disambiguation. In: KDD'2013 (2013)
18. Zhang, W., Su, J., Chen, B., Wang, W., Toh, Z., Sim, Y., Tan, C. L.: I2r-nus-msra at tac 2011: entity linking. In: Proceedings of Text Analysis Conference (TAC 2011) (2011)
19. Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., Jung, A.: Cross-lingual cross-document coreference with entity linking. In: Proceedings of Text Analysis Conference (TAC 2011) (2011)
20. Chang, A.X., Spitkovsky, V.I., Agirre, E., Manning, C.D.: Stanford-UBC entity linking at TAC-KBP, again. In: Proceedings of Text Analysis Conference (TAC 2011) (2011)
21. Radford, W., Hachey, B., Honnibal, M., Nothman, J., Curran, J.R.: Naıve but effective NIL clustering baselines–CMCRC at TAC 2011. In: Proceedings of Text Analysis Conference (TAC 2011) (2011)
22. Taylor Cassidy, Z.C., Artiles, J., Ji, H., Deng, H., Ratinov, L.A., Zheng, J., Roth, D.: CUNY-UIUC-SRI TAC-KBP2011 entity linking system description. In: Proceedings Text Analysis Conference (TAC2011) (2011)
23. Nguyen, H.T., Cao, T.H.: Named entity disambiguation: a hybrid approach. Int. J. Comput. Intell. Syst. **5**(6), 1052–1067 (2012)