A Topic-Sensitive Model for Salient Entity Linking

Lei Zhang, Cong Liu, and Achim Rettinger

Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany
{l.zhang, rettinger}@kit.edu, cong.liu@student.kit.edu

Abstract. In recent years, the amount of entities in large knowledge bases available on the Web has been increasing rapidly. Such entities can be used to bridge textual data with knowledge bases and thus help with many tasks, such as text understanding, word sense disambiguation and information retrieval. The key issue is to link the entity mentions in documents with the corresponding entities in knowledge bases, referred to as *entity linking*. In addition, for many entity-centric applications, *entity salience* for a document has become a very important factor. This raises an impending need to identify a set of salient entities that are central to the input document. In this paper, we introduce a new task of *salient entity linking* and propose a graph-based disambiguation solution, which integrates several features, especially a topic-sensitive model based on Wikipedia categories. Experimental results show that our method significantly outperforms the state-of-the-art entity linking methods in terms of precision, recall and F-measure.

1 Introduction

In recent years, large repositories of structured knowledge publicly available on the Web, such as Wikipedia, DBpedia, Freebase and YAGO, have become valuable resources for information extraction. In this regard, entity linking, which leverages such knowledge bases to link words or phrases in natural language text with the corresponding entities, has emerged as a topic of major interest.

The challenges of entity linking lie in entity recognition and disambiguation. The first stage serves to detect words or phrases in text, also called mentions, that are likely to denote entities; the second stage performs the disambiguation of the recognized mentions into entities. Many methods [1,2,3,4,5,6,7,8] have been proposed to address the problems of entity disambiguation and linking. However, these methods do not take into account the actual importance of entities w.r.t. the topics of the input document. In this work, the relation between the candidate entities and their associated categories are utilized to opt the entities that are related to the document topics.

In addition, there is an impending need to identify a set of salient entities in a document that play an important role in the content of the document, which would help to better understand its meaning or aboutness [9]. This paper focuses on the task of *salient entity linking*, especially the disambiguation of

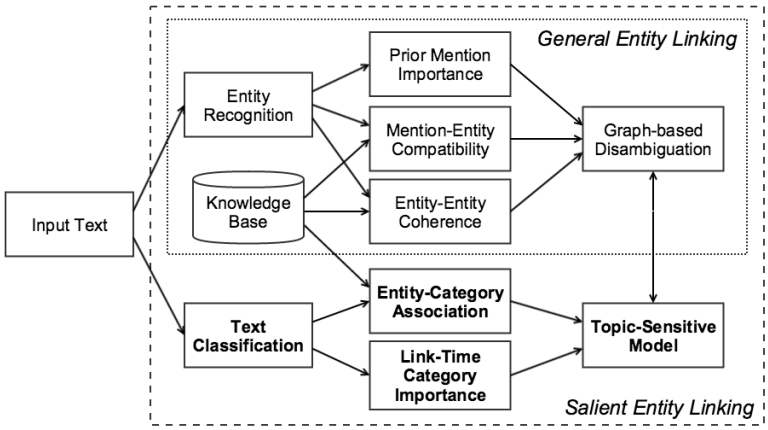

Fig. 1: Salient entity linking framework.

mentions into salient entities in a document. For this purpose, we propose a graph-based disambiguation framework, which utilizes a topic-sensitive model based on Wikipedia categories.

The rest of the paper is organized as follows. We start with an overview of our framework for salient entity linking in Sec. 2. The details of features and measures used for salient entity disambiguation are provided in Sec. 3. Based on them, we discuss the graph-based disambiguation utilizing a topic-sensitive model in Sec. 4. Evaluation results are then presented in Sec. 5, followed by the conclusions in Sec. 6.

2 Framework

Before we discuss our salient entity linking framework, we first formulate the task of entity linking and then introduce the problem of salient entity linking, an extension of the general entity linking task.

Definition 1 (Entity Linking). *Let $M = \{m_1, m_2, \ldots, m_p\}$ denote a set of entity mentions in a document D. Given a knowledge base KB containing a set of entities $E = \{e_1, e_2, \ldots, e_n\}$, the objective of* entity linking *is to determine the referent entities in KB for the mentions in M, where two functions are to be found. For entity recognition, the mentions need to be extracted from D, where a recognition function $er : D \rightarrow 2^M$ will be computed. The resulting mentions (i.e., a subset $\mu \subseteq M$) are then mapped to entities in KB, where a disambiguation function $ed : \mu \rightarrow E$ must be derived.*

Definition 2 (Salient Entity Linking). *Given a knowledge based KB and a document D, the recognition function of salient entity linking is same as general entity linking, i.e., $er : D \rightarrow 2^M$. For the set of mentions $\mu \subseteq M$ yielded by the recognition function, the disambiguation function $ed : \mu \rightarrow E \cup \{\textit{Non-Salient}\}$, which maps the set of mentions μ to entities in the KB or to non-salient entities,*

must be derived, where non-salient entities are entities with no focus of attention in D, i.e., the document D is really not about such entities.

An illustration of our salient entity linking framework consisting of several components is given in Fig. 1. In the following, we first introduce the components w.r.t. general entity linking and then discuss its extension with the components for salient entity linking by utilizing a topic-sensitive model.

For both general and salient entity linking, the input text is first processed by *entity recognition*, which detects the boundaries of mentions without knowing the actual referent entities or whether they are salient or non-salient entities. Then these mentions serve as the input of entity disambiguation, which is the focus of this work since we do not aim to compare the method's ability to recognize entity names in the input text.

Given a detected mention, its candidate referent entities are extracted from the knowledge base. For entity disambiguation regarding general entity linking, our framework combines different features including *prior mention importance*, *mention-entity compatibility* and *entity-entity coherence*. The feature of *prior mention importance* assigns the prior importance to each detected mention as weight and it will be used as the initial evidence for graph-based disambiguation. While the local feature of *mention-entity compatibility* captures the most likely entity behind the mention and the entity that best fits the context, the global feature of *entity-entity coherence* collectively captures the linked entities in a document that are related to each other. These features are then employed by *graph-based disambiguation* based on a personalized PageRank algorithm.

To aim for effective salient entity linking, we first perform *text classification* on the input text using a multi-class support vector machine (SVM) classifier based on Wikipedia categories[1] aligned with the training corpus. For each category, we compute the category probability of the input document that serves as the feature of *document-specific category importance*. In addition, we compute the strength of *entity-category association* based on the depth between each candidate entity and its categories. Such features are then incorporated into *graph-based disambiguation* using a topic-sensitive PageRank algorithm.

3 Features and Measures

In this section, we discuss the features and measures needed for salient entity disambiguation, while the graph model and algorithm will be presented in Sec. 4.

Prior Mention Importance. We employ the Wikipedia link structures for determining the *prior mention importance*. As each Wikipedia article describes an entity, article titles, redirect pages and link anchors can be used to refer to the entity. Based on the above sources, we extract all surface forms of entities.

[1] In this work, we employ the 16 second-level categories including *Mathematics*, *People*, *Science*, *Sport*, *Geography*, *Culture*, *Politics*, *Nature*, *Technology*, *Education*, *Health*, *Business*, *Belief*, *Society*, *Life* and *Concepts* in Wikipedia, where the first-level category is the fundamental category.

For each mention m with the name $m.s$ as surface form of an entity, we define the probability $P(m.s)$ that captures how likely $m.s$ refers to an entity as

$$P(m.s) = \frac{count_{link}(m.s)}{count_{link}(m.s) + count_{text}(m.s)} \tag{1}$$

where $count_{link}(m.s)$ is the number of articles that contain $m.s$ as anchor text and $count_{text}(m.s)$ is the number of articles where $m.s$ appears as raw text.

Mention-Entity Compatibility. For each mention m and its candidate referent entity e, we calculate the semantic similarity $SS(m, e)$ representing the local *mention-entity compatibility* of m and e as follows

$$SS(m, e) = \alpha \cdot LP(m, e) + \beta \cdot CS(m, e) \tag{2}$$

where $LP(m, e)$ is the link probability of e for m and $CS(m, e)$ is the context similarity between m and e, α and β are tunable parameters with $\alpha + \beta = 1$. The link probability $LP(m, e)$ can be calculated using the probability $P(e|m.s)$ capturing how likely the mention name $m.s$ refers to the entity e as follows

$$LP(m, e) = P(e|m.s) = \frac{count_{link}(e, m.s)}{\sum_{e_i \in E_{m.s}} count_{link}(e_i, m.s)} \tag{3}$$

where $count_{link}(e, m.s)$ denotes the number of links using $m.s$ as anchor text pointing to e as destination and $E_{m.s}$ is the set of entities that have the surface form $m.s$. An entity e is characterized by its textual description $e.c$, called *context* of e and a mention m is characterized by its surrounding sentences $m.c$, called *context* of m. The context similarity $CS(m, e)$ between m and e can be calculated using cosine similarity on the term vectors $e.\mathbf{c}$ of $e.c$ and $m.\mathbf{c}$ of $m.c$ as

$$CS(m, e) = cos(e.\mathbf{c}, m.\mathbf{c}) = \frac{\langle e.\mathbf{c}, m.\mathbf{c} \rangle}{|e.\mathbf{c}| \cdot |m.\mathbf{c}|} \tag{4}$$

Entity-Entity Coherence. The disambiguation is based on the feature of *entity-entity coherence*, which collectively captures the referent entities of the mentions contained in the same document that are related to each other. In this regard, we calculate the semantic relatedness between each pair of entities e_i and e_j by adopting the Wikipedia link-based measure described in [10], which is originally modeled after the Normalized Google Distance (NGD) [11], as follows

$$SR(e_i, e_j) = 1 - \frac{\log(\max(|E_i|, |E_j|)) - \log(|E_i \cap E_j|)}{\log(|E|) - \log(min(|E_i|, |E_j|))} \tag{5}$$

where E_i and E_j are the sets of entities that link to e_i and e_j in KB respectively, and E is the set of all entities in KB.

Document-specific Category Importance. For text classification of the input document, we employ John C. Platt's sequential minimal optimization for training a support vector machine (SVM) classifier [12,13]. Multi-category problems are solved using pairwise classification. To obtain proper probability

estimates, we use the option that fits logistic regression models to the outputs of the SVM classifier. In our multi-category scenario, the predicted probabilities are coupled using Hastie and Tibshirani's pairwise coupling method [14]. All these algorithms have been integrated into Weka[2], a collection of machine learning algorithms for data mining tasks. Based on that, we calculate the category probability $P(c_i)$ of the input text for each assigned category c_i, which reflects the *document-specific category importance*.

Entity-Category Association. All candidate entities are mapped to the selected Wikipedia categories. In order to measure the *entity-category association* between an entity e and its assigned category c, we define the distance $d(c, e)$ as the minimum depth at which the entity e is located in Wikipedia's category tree with the category c as the root. This is computed offline by performing a breadth-first search starting from the fundamental category that forms the root of Wikipedia's hierarchy to each entity. Then the semantic association $SA(c, e)$ between entity e and category c can be calculated as

$$SA(c, e) = \frac{1}{d(c, e)} \tag{6}$$

4 Graph Model and Algorithm

Based on the features and measures discussed in Sec. 3, we construct a directed weighted graph $G = \{N, R\}$, called *disambiguation graph*, where $N = N_M \uplus N_E \uplus N_C$ is the disjoint union of *mention* nodes N_M, *entity* nodes N_E and *category* nodes N_C, and R is the set of directed edges representing relationships between these nodes. All detected mentions and their candidate referent entities are added into N_M and N_E, respectively, while the categories that the input text belongs to are added into N_C. For each mention m and its candidate entity e, we add an edge from m to e into R. Additionally, we add an edge between e_i and e_j into R if they are connected in KB. Furthermore, for each association between an entity e and a category c, an edge from c to e will be added into R.

Once the disambiguation graph G is built, we apply a personalized PageRank algorithm [15,16] over it. The calculation of the PageRank vector Pr over G is equivalent to resolving the following equation

$$Pr = d \cdot T \cdot Pr + (1 - d) \cdot v \tag{7}$$

where T is the transition probability matrix, v is the initial evidence vector and d is the so called damping factor, usually set as 0.85. Each entry T_{ij} in T is the evidence propagation ratio from node i to node j, which is computed in Eq. 8.

$$T_{ij} = \begin{cases} \frac{SS(m_i, e_j)}{\sum_{k \in N_E(i)} SS(m_i, e_k)} & \text{if } i \in N_M,\ j \in N_E \\ \frac{SR(e_i, e_j)}{\sum_{k \in N_E(i)} SR(e_i, e_k)} & \text{if } i \in N_E,\ j \in N_E \\ \frac{SA(c_i, e_j)}{\sum_{k \in N_E(i)} SA(c_i, e_k)} & \text{if } i \in N_C,\ j \in N_E \end{cases} \tag{8}$$

[2] `http://www.cs.waikato.ac.nz/ml/weka`

where $N_E(i)$ is the set of entity nodes such that for each node $k \in N_E(i)$, there is an edge from i to k in G. The entry v_i in v is the initial evidence representing the prior importance of a mention m_i if $i \in N_M$ or the document-specific importance of an category c_i if $i \in N_C$, which is calculated as follows

$$v_i = \begin{cases} \frac{\lambda \cdot P(m_i)}{\lambda \cdot \sum_{k \in N_M} P(m_k) + \eta \cdot \sum_{k \in N_C} P(c_k)} & \text{if } i \in N_M \\ \frac{\eta \cdot P(c_i)}{\lambda \cdot \sum_{k \in N_M} P(m_k) + \eta \cdot \sum_{k \in N_C} P(c_k)} & \text{if } i \in N_C \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where λ and η are tunable parameters with $\lambda + \eta = 1$, which reflect the sensitivity of prior mention importance and document-specific category importance to the final probability of each candidate entity. When $\eta = 0$, our method reduces to general entity linking without considering the topic-sensitive model. In contrast, when $\lambda = 0$, the initial evidence of the graph-based disambiguation only depends on the category importance.

As a result of the personalized PageRank algorithm, each candidate entity e receives a final probability $P(e)$. For each mention m having a set of candidate entities E_m, we choose the entity with the maximal probability as the predicted linking entity, i.e., $e_m = \arg\max_{e \in E_m} P(e)$. The process discussed above doesn't distinguish between salient and non-salient entities. In order to deal with salient entity linking, one important task of the *topic-sensitive model* is to validate whether the predicted linking entity e_m for mention m is a salient entity. For this purpose, we learn a threshold τ such that if $P(e_m)$ is greater than τ we return e_m as the linking entity for m, otherwise we return *Non-Salient*.

5 Experiments

We now discuss the experiments we performed to assess the performance of our approach. As the knowledge base, we used the English Wikipedia snapshot from July 2013. We employed the Reuters-128 entity salience dataset[3], which is an extension of a part of the N3 entity linking datasets [17]. The Reuters-128 dataset is an English corpus and it contains 128 economic news articles. The dataset contains information for 880 named entities with their position in the document and a URI of a DBpedia resource identifying each entity. The salience dataset extends the Reuters-128 dataset also with 3,551 common entities.

In order to construct the dataset, entity salience information was obtained by crowdsourcing salience information using the CrowdFlower platform. For each named and common entity in the Reuters-128 dataset, the authors of the dataset collected at least three judgements. Only judgments from annotator with trust score higher than 70% were considered as trusted judgements. If the trust score of an annotator falls bellow 70%, all his/her judgements were disregarded. Finally, each named and common entity in the dataset has been classified in one of the following classes[4]:

[3] `https://github.com/KIZI/ner-eval-collection`

[4] `http://ner.vse.cz/datasets/entitysalience-collection`

Methods	Mic. Prec.	Mic. Rec.	Mic. F1	Mac. Prec.	Mac. Rec.	Mac. F1.
DBpedia Spotlight [2]	0.45	0.39	0.41	0.45	0.37	0.40
Wikipedia Miner [1]	0.60	0.48	0.54	0.60	0.42	0.52
NERD-ML [5,7]	0.67	0.50	0.57	0.65	0.46	0.54
WAT [4,8]	0.35	0.32	0.34	0.36	0.33	0.34
AGDISTIS [6]	0.73	0.50	0.59	0.73	0.48	0.58
Our Method (General)	0.70	0.46	0.56	0.69	0.45	0.55
Our Method (Salient)	**0.83**	**0.51**	**0.63**	**0.82**	**0.50**	**0.62**

Table 1: The experimental results.

- *Most Salient* - Entities with the highest focus of attention in the article. The document is mostly about the these entities, or the entities play a prominent role in the content of the article.
- *Less Salient* - Entities with less focus of attention in the article. The entities play an important role in some parts of the content of the article.
- *Not Salient* - The article is really not about the entities

In our experiments, we consider the entities in both classes *Most Salient* and *Less Salient* as salient entities, while entitites belonging to *Not Salient* are considered as non-salient entities. Using the Reuters-128 entity salience dataset, we conducted the experiments to compare our approach with several entity linking methods. We used two variants of our approach, one employs only the graph-based disambiguation for general entity linking ($\lambda = 1$ and $\eta = 0$) and the other integrates the topic-sensitive model with the goal of salient entity linking ($\lambda = 0.2$ and $\eta = 0.8$). All the methods should label each mention with either the correct entity or *Not Salient*. Note that we restrict the input to the labeled mentions to compare the method's ability to distinguish between salient entity and non-salient entity, not its ability to recognize entity names in the input text. The adopted evaluation criteria include Micro-Precision, Micro-Recall, Micro-F1, Macro-Precision, Macro-Recall and Macro-F1.

The experimental results are shown in Table 1. By utilizing the topic-sensitive model, our approach to salient entity disambiguation significantly outperforms the baselines in terms of all evaluation criteria. Regarding the two variants of our approach, it clearly shows that the topic-sensitive model indeed contributes to the final performance improvement.

6 Conclusions

In this paper, we introduce the task of salient entity linking that existing entity linking solutions cannot well address. For tackling this new problem, we propose a graph-based disambiguation framework, which integrates several features including prior mention importance, mention-entity compatibility, entity-entity coherence and in particular a topic-sensitive model capturing entity-category association and document-specific category importance. We have experimentally shown that our approach achieves a significant improvement over the baselines. The evaluation results also show that the topic-sensitive model indeed helps with the salient entity disambiguation.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

References

1. Milne, D.N., Witten, I.H.: Learning to link with wikipedia. In: CIKM. (2008) 509–518
2. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: I-SEMANTICS. (2011) 1–8
3. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: SIGIR. (2011) 765–774
4. Ferragina, P., Scaiella, U.: Fast and accurate annotation of short texts with wikipedia pages. IEEE Software **29**(1) (2012) 70–75
5. van Erp, M., Rizzo, G., Troncy, R.: Learning with the web: Spotting named entities on the intersection of NERD and machine learning. In: #MSM. (2013) 27–30
6. Usbeck, R., Ngomo, A.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: AGDISTIS - graph-based disambiguation of named entities using linked data. In: ISWC. (2014) 457–471
7. Rizzo, G., van Erp, M., Troncy, R.: Benchmarking the extraction and disambiguation of named entities on the semantic web. In: LREC. (2014) 4593–4600
8. Piccinno, F., Ferragina, P.: From tagme to WAT: a new entity annotator. In: ERD@SIGIR. (2014) 55–62
9. Gamon, M., Yano, T., Song, X., Apacible, J., Pantel, P.: Identifying salient entities in web pages. In: CIKM. (2013) 2375–2380
10. Witten, I., Milne, D.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: WIKIAI. (2008) 25–30
11. Cilibrasi, R., Vitányi, P.M.B.: The google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3) (2007) 370–383
12. Platt, J.C.: Advances in kernel methods. MIT Press, Cambridge, MA, USA (1999) 185–208
13. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to platt's smo algorithm for svm classifier design. Neural Comput. **13**(3) (2001) 637–649
14. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: NIPS. (1997) 507–513
15. Jeh, G., Widom, J.: Scaling personalized web search. In: WWW. (2003) 271–279
16. Haveliwala, T.H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans. Knowl. Data Eng. **15**(4) (2003) 784–796
17. Röder, M., Usbeck, R., Hellmann, S., Gerber, D., Both, A.: N^3 - A collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In: LREC. (2014) 3529–3533