Automatic Selection of Reference Pages in Wikipedia for Improving Targeted Entities Disambiguation

Takuya Makino
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Japan
makino.takuya@jp.fujitsu.com

Abstract

In Targeted Entity Disambiguation setting, we take (i) a set of entity names which belong to the same domain (target entities), (ii) candidate mentions of the given entities which are texts that contain the target entities as input, and then determine which ones are true mentions of "target entity". For example, given the names of IT companies, including Apple, we determine Apple in a mention denotes an IT company or not. Prior work proposed a graph based model. This model ranks all candidate mentions based on scores which denote the degree of relevancy to target entities. Furthermore, this graph based model could utilize reference pages of target entities. However, human annotators must select reference pages in advance. We propose an automatic method that can select reference pages. We formalize the selection problem of reference pages as an Integer Linear Programming problem. We show that our model works as well as the prior work that manually selected reference pages.

1 Introduction

The enterprise is typically interested in customer's opinions. One of the methods to analyze customer's opinions is to collect mentions which contain product names. We would get a noisy mention collection if we use a simple method which extracts mentions that contain product names, since the product names may be used as other meanings.

Wang et al. (2012) proposed a new task which they referred to as Targeted Entity Disambiguation (TED). In this problem setting, we take (i) a set of entity names which belong to the same domain (target entities), (ii) candidate mentions of the given entities which are texts that contain the target entity entities as input, and then determine which ones are true mentions for the target entities. TED is different from traditional Word Sense Disambiguation or Entity Linking. Word Sense Disambiguation can be viewed as a classification task in which word senses are the classes (Navigli, 2009) and Entity Linking is the task of linking name in Web text with entities in Wikipedia (Han et al., 2011). The uniqueness of this problem is that the entities are all in the same domain (referred to as the target domain) and not necessarily included in a knowledge base such as DBpedia, Freebase or YAGO.

Wang et al. (2012) realized TED with a graph based model. In their graph based method, a target entity in a mention is regarded as a node, and the weight of an edge is determined according to context similarity, and a prior score of node that is determined according to the unique number of target entities in the mention. This graph is called as a mention graph. Using mention graph, the authority of each mention is calculated with MentionRank which is a variant of PageRank (Page et al., 1999). This authority denotes a score of how likely this node is in the target domain. In addition, MentionRank could integrate external knowledge such as Wikipedia. For each target entity, a reference page is added as a virtual node to the graph. Since reference pages can be regarded as true mentions, the prior scores of virtual nodes are higher than other mentions. This extended method can propagate the score of the virtual node of each entity to candidate mentions which are likely true. Although the use of reference pages works well, human annotators must select these reference pages.

In Word Sense Disambiguation and Entity Linking, there are some collective approaches (Hoffart et al., 2011; Kulkarni et al., 2009). In this paper, we apply this technique to the selection problem of reference pages for TED. To select refer-

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 106–110,
Gothenburg, Sweden, April 26-30 2014.

ence pages, we collect candidate reference pages of target entities from Wikipedia in advance. If the name of a target entity has a disambiguation page in Wikipedia, we have two or more candidate reference pages. Then we formalize the problem of reference page selection as an Integer Linear Programming problem. Our model is going to maximize the summation of similarities between selected pages under some constraints. Thus, coherent pages are selected as reference pages. Our method does not require any knowledge except for names of target entities. We give only target entities as input to select reference pages. Our method shows competitive accuracy of the prior method with manually selected reference pages.

2 Task Definition

Following previous work, we assume that all occurrences of a name in a mention refer to the same entity (e.g., occurrences of the string "Apple" in a single mention either all refer to the IT company or all refer to the fruit) (Wang et al., 2012).

TED is defined as follows.

Definition 1 (Targeted Entity Disambiguation)**.** *Given input of a target entity set* $E = \{e_1, ..., e_n\}$*, a mention set* $D = \{d_1, ..., d_n\}$ *and candidate mentions* $R = \{(e_i, d_j)|e_i \in E, d_j \in D\}$*, output score* $r_{ij} \in [0, 1]$ *for every candidate mention* $(e_i, d_j) \in R$*.*

3 Related Work

Wang et al. (2012) proposed MentionRank to address TED. MentionRank is similar to PageRank. This model is based on three hypotheses:

1. **Context similarity:** The true mentions across all the entities, across all the mentions will have more similar contexts than the false mentions of different entities.

2. **Co-Mention:** If multiple target entities are co-mentioned in a mention, they are likely to be true mentions.

3. **Interdependency:** If one or more mentions among the ones with similar context is deemed likely to be a true mention, they are all likely to be true mentions.

In a mention graph, a node (e_i, d_j) denotes an entity e_i in mention d_j. The weight of edge between (e_i, d_j) and (e'_i, d'_j) is denoted as $w_{ij,i'j'}$ which is a variable normalized by context similarity $\mu_{ij,i'j'}$. Context similarities are normalized to avoid "false-boost" problem. "false-boost" problem is boosting ranking score of false mentions in a false mentions group. The normalized weight of the edge is defined as follows:

$$w_{ij,i'j'} = \begin{cases} \frac{z_{ij}}{k} & \text{if} \quad i = i', \\ \frac{\mu_{i'j',ij}}{V_i Z} + \frac{z_{ij}}{k} & \text{otherwise.} \end{cases} \tag{1}$$

$$z_{ij} = 1 - \frac{\sum_{i' \neq i} \sum_{j'} \mu_{i'j',ij}}{V_i Z}, \tag{2}$$

$$Z = \max_{i,j} \frac{\sum_{i' \neq i} \sum_{j'} \mu_{i'j',ij}}{V_i}, \tag{3}$$

where, V_i denotes the number of candidate mentions that contain e_i (i.e. $V_i = |\{d_j|(e_i, d_j) \in R\}|$). k denotes the number of all candidate mentions (i.e. $k = |R|$). Co-mention is represented by a prior score. Wang et al. (2012) defined prior score π_{ij} of (e_i, d_j) as the number of unique names of target entities occurred in d_j.

The final score of each mention is decided by its prior score estimation as well as the score of the other correlated mentions.

$$r_{ij} = \lambda p_{ij} + (1 - \lambda) \sum_{i',j'} w_{ij,i'j'} r_{i'j'}, \tag{4}$$

where λ is the dumping factor. p_{ij} denotes prior score of (e_i, d_j): $p_{ij} = \pi_{ij} / \sum_{i',j'} \pi_{i'j'}$

Although this model works even if only the names of entities are given as input, we can extend this model to integrate external knowledge such as Wikipedia. For example, we can add reference pages for each entity as virtual nodes. Since we can assume that the reference page of a target entity is a true mention with a high confidence, we assign a high prior score than the other mentions. This causes the group of candidate mentions which have similar contexts with the reference pages to get higher scores. One example of using reference pages is to add a set of reference pages $\{a_i|1 \leq i \leq n\}$ into the mention graph. a_i denotes the reference page of entity e_i.

4 Proposed Method

In this section, we propose our approach for automatic selection of reference pages. In the domain of Word Sense Disambiguation and Entity Linking, some researches proposed the methods which

Figure 1: Article "Apple (disambiguation)" in Wikipedia

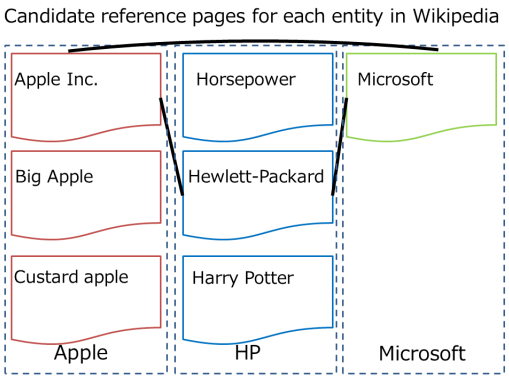

Figure 2: Automatic selection of reference pages from disambiguation pages in Wikipedia: selected pages contains same categories or similar contents (They are connected by edge).

are based on coherence between mentions (Hoffart et al., 2011; Kulkarni et al., 2009; Han et al., 2011). Our method does not require any knowledge except for the names of target entities. We give only target entities as input. Target entities in Wikipedia have two characteristics.

- A name of an ambiguous target entity tends to have a disambiguation page.
- The articles that are in the same domain have the same categories or contain similar contents.

In Wikipedia, there are disambiguation pages like Figure 1. "Apple (disambiguation)" contains apple as a plant, an IT company, a music album, and so on. To collect candidate reference pages, we use these disambiguation pages.

Kulkarni et al. (2009) formalized entity linking as an Integer Linear Programming problem and then relaxed it as a Linear Programming problem. They considered a coherence score which takes higher value if the selected articles have similar contents. Their framework can be used for entity linking and word sense disambiguation. In this paper, we use this coherence score to select reference pages. We show an image of an automatic selection of reference pages in Figure 2. In Figure 2, the target entities are Apple, HP and Microsoft. Although we have only one page for Microsoft, we have two or more candidate reference pages, since Apple and HP have disambiguation pages. Then we need to select reference pages for Apple and HP. If the name of a target entity is not in Wikipedia, we have no reference page for that target entity. The goal of this example is to select "Apple Inc." for Apple and "Hewlett-Packard" for HP (Selecting "Microsoft" for Microsoft is trivial). We regard these selected articles as reference pages for target entities.

We assume that the number of true reference page a_i for target entity e_i is one and select one reference page for each target entity. For each target entity, we select articles which the have same categories or similar contents from the set of candidate reference pages $\{c_{ik}|1 \leq k \leq l\}$ since we assume that the articles in the same domain have the same categories or contain similar contents. In fact, our model is going to maximize the summation of similarities between selected pages under some constraints. We formalize this selection as follows:

$$max. \quad \sum_{i,k} \sum_{i',k'} e_{ik,i'k'} x_{ik,i'k'},$$

$$s.t. \quad \forall i, \ \sum_{k} y_{ik} = 1, \tag{5}$$

$$y_{ik} \geq x_{ik,i'k'}; \ \forall i,k,i',k', \tag{6}$$

$$y_{i'k'} \geq x_{ik,i'k'}; \ \forall i,k,i',k', \tag{7}$$

$$x_{ik,i'k'} \in \{0,1\}; \ \forall i,k,i',k', \tag{8}$$

$$y_{ik} \in \{0,1\}; \ \forall i,k, \tag{9}$$

$e_{ik,i'k'}$ denotes the weight of the edge between candidate reference pages c_{ik} and $c_{i'k'}$. $x_{ik,i'k'}$ takes 1 if c_{ik} is selected, 0 otherwise. y_{ik} takes 1 if the edge between c_{ik} and $c_{i'k'}$ is selected, 0

	n	k	#cand	%Positive
Car	21	1809	21.5	29.9
Magazine	28	2741	17.9	43.5

Table 1: Datasets: n is # of entities, k is # of candidate mentions, #cand is average # of candidate reference pages for each entity and %Positive is % of true mentions in all candidate mentions

n=5	Car	Magazine
MentionRank	39.74	61.07
MentionRank+manVN	39.14	70.94†
MentionRank+randomVN	37.85†	65.01
Proposed method	44.21	65.86
n=10		
MentionRank	49.23	65.90†
MentionRank+manVN	47.21†	70.85
MentionRank+randomVN	45.13†	68.38
Proposed method	50.84	69.81
n=15		
MentionRank	46.50†	65.77†
MentionRank+manVN	44.29	69.38
MentionRank+randomVN	39.21†	67.89
Proposed method	42.77	69.02

Table 2: Mean average precision for each dataset

otherwise. Constraint (5) ensures that always one article is selected for each entity. Constraints (6) and (7) ensure that when $x_{ik,i'k'} = 1$, y_{ik} and $y_{i'k'}$. In this paper, we defined $e_{ik,i'k'}$ as cosine similarity of two vectors of words those weights are tfidf.

5 Experiments

We used weblogs written in Japanese for experiments. Following the previous work, we created two datasets: Car and Magazine. A summary of each dataset is shown in Table 1.

- Car: Target entities include car names such as Prius and Harrier.
- Magazine: Target entities include magazine names such as MORE and LEE.

We randomly selected 5, 10 or 15 entities from each target entities for 10 times and conducted experiment for each dataset with parameter λ = 0.15. We conducted significance test using Wilcoxon signed-rank test. Table 2 lists the experimental results on these datasets. In Table 2, MentionRank+manVN denotes MentionRank with virtual nodes that are selected manually (Wang et al., 2012). MentionRank+randomVN denotes MentionRank with virtual nodes that are selected randomly from candidate reference pages in Wikipedia. Proposed method denotes the MentionRank with virtual nodes that are selected automatically using ILP. Values with †in Table 2 indicate that there are significant differences between mean average precision of proposed method and the others. Five results of proposed methods are better than those of MentionRank, there are significant differences on two results. Furthermore, all the results of proposed method is better than those of MentionRank+randomVN and there are significant differences on three results. Four results of proposed method is worse than those of MentionRank+manVN, however there is a significant difference on only one of those results. From these results, we can see that use of reference pages automatically selected by our method improves mean average precision. In Magazine, several entities are not ambiguous and we could get true reference pages easily. Therefore, we think proposed method did not show any significant differences compared with MentionRank+randomVN. Also, in Car, several entities are not ambiguous but these reference pages belong to domains other than Car domain. As a result, we think that some results are worse than MentionRank. For example, entity "86" which is a kind of car have only one reference page that belongs to number domain.

6 Conclusion

In this paper, we proposed an automatic selection method of reference pages for Target Entity Disambiguation. Our method that uses automatically selected reference pages showed better performance than MentionRank without reference pages and competitive mean average precision with MentionRank with manually selected reference pages.

Since our framework always selects one reference page for each target entity even if a reference page does not exist in Wikipedia or one or more reference pages exist in Wikipedia, we need to refine our framework in future work. An another improvement would be to assign prior scores for virtual nodes according to coherence score between the other virtual nodes.

References

Xianpei Han, Le Sun, and Jun Zhao. 2011. 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*, SIGIR '11, pages 765–774, New York, NY, USA. ACM.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 457–466, New York, NY, USA. ACM.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November.

Chi Wang, Kaushik Chakrabarti, Tao Cheng, and Surajit Chaudhuri. 2012. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 719–728, New York, NY, USA. ACM.