Typicality-focused Comparative Analysis of Entities

Yijun DUAN† and Adam JATOWT†

† Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{yijun,adam}@dl.kuis.kyoto-u.ac.jp

Abstract Comparison is an effective strategy extensively adopted in practice. A natural prerequisite step of comparison is to find comparable entities. In this paper, we propose a novel summarization system to address the task of automatically generating typical comparable entity pairs given two comparable collection of entities. The system first apply the idea of typicality analysis to measure the representativeness of each entity. Then, the system generate a diverse set of typical comparables based on a concise integer linear programming framework. We then experimentally demonstrate the effectiveness of our system on several Wikipedia categories.

Key words Comparable Entity Identification, Typicality Analysis, Integer Linear Programming

1. Introduction

Comparison is an effective strategy extensively adopted in practice for people to discover the commonality and difference between two or more objects. People can benefit from such information for needs such as analyzing trends, gain insights about similar situations, making a better decision and so on. A natural prerequisite step of comparison is to find comparable entities. What is more, learning from examples is regarded an effective strategy extensively adopted in the daily life. Good examples are often more easy to understand than high-level feature descriptions in learning concepts or categories of entities. Therefore, given two comparable collection of entities (e.g. list of Japanese cities and United States cities), it would be very useful to automatically find a diverse set of comparable entity pairs (e.g. "Tokyo" with "Los Angeles", "Okinawa" with "Hawaii") for such pairs not only reveal contrastive knowledge, but also are quite understandable.

In this study, we propose that comparable entities are entities which have proper comparative aspect (i.e., to what the entities are compared). For instance, "Tokyo" and "Beijing" are comparable for they both serve as the capital of an Asian country. Note that such aspects can be mentioned explicitly in text, or be indicated implicitly by semantically similar context of entities. For example, a comparative sentence *"iPod is superior to cassette in terms of portability"* can clearly describe the high comparability between "iPod" and "cassette", while the following text

Cleveland Cavaliers won the 2016 NBA finals.

Golden State Warriors Won 2017 NBA Championship. also contain the evidences of proper comparison between "Cleveland Cavaliers" and "Golden State Warriors". Explicit comparative text are useful but they do not appear frequently thus show a limited coverage. To solve this issue, we consider the comparability between two entities as the similarities of their context. Furthermore, we apply distributed word embedding technique in order to obtain the context vector of each entity [12] [13]. Thus the comparability of two entities (e_A and (e_B is computed by:

$$Comp(e_A, e_B) = Sim_{cosine}(\omega(e_A), \omega(e_B)) \qquad (1)$$

where $\omega(e_A)$ and $\omega(e_B)$ are the context vector of entities (e_A and (e_B, respectively.

The problem of detecting comparable entities is however not trivial resulting from the following reasons: (1) the input collections of entities can be very large and may cover a great amount of latent diverse subgroups. Naturally, only typical entities should be chosen for comparing. For instance, to compare mammals with other animals, typical examples of mammals such as lions should be preferred rather than using atypical instances like platypuses (which lay eggs instead of giving birth to live young). This is because typical instances are usually associated with more representative features and thus are less likely to cause misleading. However it is difficult to estimate entity typicality appropriately over a broad and diverse set of entities. (2) Given the limitation on the size of output, selecting an optimal subset of both typical and comparable entities pairs is a very challenging problem for they should best reveal the overall representativeness and comparability of selected entities.

To conquer the challenges, we propose a novel summarization system to address the task of generating typical comparables. First of all, we formulate the measurement of entity typicality inspired by previous research in psychology and cognitive science. To make an entity typical as a whole in a base set comprising diverse subgroups, it should be quite representative in a fairly significant member group. Secondly, inspired by the popular Affinity Propagation algorithm, we propose a concise integer linear programming framework which detects typical entities (which we call exemplars) and generates comparable pairs from detected exemplars simultaneously. Based on this formulation, the exactly optimal solution can be obtained and validated.

To sum up, we make the following contributions: (1) We introduce a new research problem of automatically discovering typical comparable entity pairs from two comparable collections of entities. (2) We then develop a novel summarization system to address this task based on an effective entity typicality measurement and a concise integer linear programming framework. (3) we perform extensive experiments on several Wikipedia categories, which prove the effectiveness of our approach.

The reminder of this paper is organized as follows. We first survey the related work in Section 2. We formulate our research problem in Section 3. Section 4 introduces our proposed entity typicality measurement. The ILP formulation for generating typical comparable entity pairs is presented in Sections 5. We describe the experimental setup and experiment results in Section 6. We conclude the paper and outline the future work in the last section.

2. Related Work

2. 1 Comparable Entities Mining

The task of comparable entity mining has attracts much attention in the NLP and Web mining communities [5] [2] [6] [3] [4] [1]. Approaches to this task include hand-crafted extraction rules [8], supervised machine-learning methods [9] [10] and weakly-supervised methods [6] [2]. Jindal *et al.* is the first proposing a typical two-step system in finding comparable entities which first tackles a classification problem (i.e., whether a sentence is comparative) and then a label problem (i.e., which part of the sentence is the desideratum) [3] [4]. Some following work refines the system by using a bootstrapping algorithm [6], or extends the idea of mining comparables to different corpus including query logs [1] [2] and comparative questions [6]. In addition, comparable entities mining is also strongly related to the problem of automatic structured information extraction, comparative summarization and named entity recognition. Some work lies in the intersection of these tasks [7] [11].

How is our study related?

To the best of our knowledge, we are the first to focus on the typicality of extracted comparable entities using the idea of typicality analysis from psychology.

2. 2 Typicality Analysis

Typicality of entities has been widely discussed in the field of psychology [29], [30]. Typical entities are usually judged as "better examples" of a category. There are generally two types of determinants of such category representativeness. One determinant is called the central tendency [27], which is either one or several existing or imaginary very representative entity(entities). The typicality of an entity is determined by its similarities to the central tendency. The other determinant is the stimulus similarity [28]. The more similar an instance is to the other members of its category, and the less similar it is to members of the contrast categories, the higher the typicality rating it has.

Besides, typicality is a concept often modeled using any of the prototype view [24], the exemplar view [25] and the schema view [26]. These theories have different prospectives in the form of the central tendency of the associated category. The prototype view proposes using abstract prototypes to represent a category instead of entities existing in real life suggested by the example view, while the scheme views models categories with artificial intelligence knowledge representation. In the recent years, the idea of typicality has also been introduced in areas such ontology design [21], and query answering [19] [20].

How is our study related?

In this study, we propose a specifically designed typicality measure for our comparable entity identification (CEI) task. Our typicality measure is in the general spirit of typicality measures in psychology. However, we improve them by computing the exact answer using a concise integer linear programming formulation.

2. 3 Affinity Propagation

The Affinity Propagation (AP) algorithm [14] is a clustering algorithm which has been prove useful in many scenarios such as computer vision and computational biology tasks [15] [17]. AP views the clustering as identifying a subset of exemplars and assigns each non-exemplar item to an exemplar item. It takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. AP embodies the characteristics of good robustness and high accuracy. However, it does not guarantee to find the optimal set of exemplars.

How is our study related?

In spired by AP, we formulate the CEI as a process of identifying a subset of typical comparable entity pairs. It has been empirically found that using AP for solving a special case of objective (Eq. (4)) suffers considerably from convergence issues [16]. Thus, we propose a concise integer linear programming (ILP) formulation for solving CEI, and we use the bound-and-branch method to obtain the optimal solution.

3. Problem Definition

In this study, our task is to briefly sum up the commonalities and differences between two comparable topics by using comparable entity pairs. The summarization system is given two collections of entities, each of which is related to a topic. The system aims to find latent comparative aspects, and generate entities from the two topics of those aspects in a pairwise way.

Formally, let two comparable topics denoted by T_A and T_B, and D_A and D_B be the corresponding collections of entities respectively, the task is to discover m comparable entity pairs $[p_1, p_2, ..., p_m]$ to form a concise summary conveying the most import comparisons, where $p_i = (e_i^A, e_i^B)$. e_i^A and e_i^B are entities from D_A and D_B in the same latent aspect respectively. The pairs should have good quality, i.e., each entity should be representative in its topic, and entities within the same pair should be fairly comparable. Moreover, selected entities should cover as many subgroups as possible to avoid redundancy and reflect intrinsic diversity.

4. Estimation of Entity Typicality

Learning from examples is an effective strategy extensively adopted in learning, and good examples should be typical. Typicality analysis has been widely studied in psychology and cognitive science. In this study, we apply the strategy of using typical examples to summarize and analyze a large set of entities to discovering comparable entity pairs. We define the typicality of an entity e with regard to a set of entities S as $Typ(e, S)$. Intuitively, entities to be selected for comparison should be typical in their categories, namely $Typ(e_i^A, D_A)$ and $Typ(e_i^B, D_B)$ should be as high as possible, where $p_i = (e_i^A, e_i^B)$ is an expected entity pair.

Suggested by the previous research in typicality analysis, an entity e in a set of entities S is more typical than the others if e is more likely to appear in S. We denote the likelihood of an entity e given a set of entities S by $L(e|S)$. However, it is not appropriate to simply use $L(e|S)$ as an estimator of typicality $Typ(e, S)$ considering the characteristics of our task. First of all, the collections of entities for comparison can be very large, thus may cover a great amount of types of entities. For example, if we want to compare Japanese scientists and United States scientists, each collection will cover multiple types of entities such as mathematician, physicist, chemist and so on. It is very difficult for a single entity to represent all of them, and an entity can only represent entities in the same or similar types well. In addition, different entity types vary in significance. For instance, "politicians" are far more important than "portable music players" in a newspaper corpus. Naturally, entities typical in a salient entity type should be more important than those belonging to trivial types.

Given a base set S_A including k latent subgroups $[S_1^A, S_2^A, ..., S_k^A]$ and the contrast set S_B, let e_i^t denote the ith entity in the tth subgroup of S_A. To make e_i^t typical as a whole, we state two hypotheses:

Hypothesis 1 *e_i^t should be fairly representative in S_t^A.*

Hypothesis 2 *The significance of S_t^A should be fairly large.*

The typicality of e_i^t with respect to S_A is defined as follows:

$$Typ(e_i^t, S^A) = L(e_i^t|S_t^A) \cdot \frac{|S_t^A|}{|S^A|} \tag{2}$$

where $L(e_i^t|S_t^A)$ measures the representativeness of e_i^t with regard to S_t^A. e_i^t is typical and discriminative in S^A if the difference between its representativeness in S^A and S^B is large. In addition, $\frac{|S_t^A|}{|S^A|}$ depicts the size of S_t^A, which can be regarded as an estimator of significance. e_i^t is typical if the number of entities in its group is very large.

The likelihood of an entity e given a set of entities S by $L(e|S)$ is the posterior probability of e given S, which can be computed using probability density estimation methods. Many model estimation techniques have been proposed including parametric and non-parametric density estimation. In this study, we use kernel estimation [23], for it does not require a certain distribution assumption and can estimate unknown data distributions effectively. Moreover, we choose the commonly used Gaussian kernels. We set the bandwidth of the Gaussian kernel estimator $h = \frac{1.06s}{\sqrt[5]{n}}$ as suggested in [22], where n is the size of the data and s is the standard deviation of the data set. Formally, given a set of entities $S = (e_1, e_2, ..., e_n)$, the underlying likelihood function is approximated as:

$$L(e|S) = \frac{1}{n}\sum_{i=1}^{n} G_h(e, e_i) = \frac{1}{n\sqrt{2\pi}} e^{-\frac{d(e,e_i)^2}{2h^2}} \tag{3}$$

where $d(e, e_i)$ is the distance between e and $, e_i$, and $G_h(e, e_i)$ is a *Gaussian kernel*.

5. ILP Formulation for CEI

In this section, we describe our proposed method for discovering comparable entity pairs. Given two collections of articles D_A and D_B expressing comparable topics T_A and

T_B respectively, the output of the summarization system are supposed to be m comparable entity pairs $[p_1, p_2, ..., p_m]$, where each pair composes of two entities from D_A and D_B in the same latent comparative aspect respectively. Intuitively, if the selected entities are very typical in their collections (i.e. representative in their own underlying subgroups), and entities within the same pair are fairly comparable (i.e. their context vector are similar), the generated pairs will highlight the commonalities and differences between documents compared and be of high quality.

Therefore, we develop an ILP formulation which detects typical entities (which we call exemplars) and generates comparable pairs from detected exemplars simultaneously. More explicitly, we formulate the task as a process of selecting a subset of k exemplars for each topic and ranking m entity pairs based on the identified exemplars. Each non-exemplar entity is assigned to an exemplar item based on some measure of similarity, and each exemplar e represents a subgroup comprised of all non-exemplar entities assigned to e . This motivation follows the aforesaid AP algorithm. On the one hand, we wish to maximize the overall typicality of selected exemplars w.r.t. there representing groups. On the other hand, we expect to maximize the overall comparability of the top m entity pairs, where each pair consisting of two exemplars from different topics.

We then introduce some notations used in our propose method. Let e_i^A denotes the ith entity in D_A. $M_A = [m_{ij}]^A$ is a $n_A \times n_A$ binary square matrix such that n_A is the number of entities within D_A. m_{ii}^A indicates whether entity e_i^A is selected as an exemplar or not, and $m_{ij:i \neq j}^A$ represents whether entity e_i^A votes for entity e_j^A as its exemplar. Similar to M_A, the $n_B \times n_B$ binary square matrix M_B depicts how entities belonging to D_B choose their exemplars, where n_B is the number of entities within D_B. m_{ii}^B indicates whether entity e_i^B is selected as an exemplar or not, and $m_{ij:i \neq j}^B$ represents whether entity e_i^B votes for entity e_j^B as its exemplar. Different from M_A and M_B, $M_T = [m_{ij}]^T$ is a $n_A \times n_B$ binary matrix whose entry m_{ij}^T denotes whether entities e_i^A and e_j^B are paired together as the final result. Then the following ILP problem is designed for the task of selecting k exemplars for each topic and ranking m comparable entity pairs.

$$max \lambda \cdot m \cdot \left[T'(M_A) + T'(M_B) \right] + (1-\lambda) \cdot 2k \cdot C'(M_D) \quad (4)$$

$$T'(M_X) = \sum_{i=1}^{n_X} m_{ii}^X \cdot Typ(e_i^X, G(e_i^X)), X \in \{A, B\} \quad (5)$$

$$C'(M_T) = \sum_{i=1}^{n_T} \sum_{j=1}^{n_T} m_{ij}^T \cdot Comp(e_i^A, e_i^B) \quad (6)$$

$$G(e_i^X) = \left\{ e_j^X | m_{ji}^X = 1 \right\}, i \in \{1, ..., n_X\}, j \in \{1, ..., n_X\}, X \in \{A, B\} \quad (7)$$

$$s.t. \quad m_{ij}^X \in 0, 1, i \in \{1, ..., n_X\}, j \in \{1, ..., n_X\}, X \in \{A, B, T\} \quad (8)$$

$$\sum_{i=1}^{n_X} m_{ii}^X = k, X \in \{A, B\} \quad (9)$$

$$\sum_{j=1}^{n_X} m_{ij}^X = 1, i \in \{1, ..., n_X\}, X \in \{A, B\} \quad (10)$$

$$m_{jj}^X - m_{ij}^X \geqq 0, i \in \{1, ..., n_X\}, j \in \{1, ..., n_X\}, X \in \{A, B\} \quad (11)$$

$$\sum_{i=1}^{n_T} \sum_{j=1}^{n_T} m_{ij}^T = m \quad (12)$$

$$m_{ii}^A - m_{ij}^T \geqq 0, i \in \{1, ..., n_A\}, j \in \{1, ..., n_B\} \quad (13)$$

$$m_{jj}^B - m_{ij}^T \geqq 0, i \in \{1, ..., n_B\}, j \in \{1, ..., n_A\} \quad (14)$$

$$\sum_{j=1}^{n_B} m_{ij}^T \leqq 1, i \in \{1, ..., n_A\} \quad (15)$$

$$\sum_{i=1}^{n_A} m_{ij}^T \leqq 1, j \in \{1, ..., n_B\} \quad (16)$$

We now explain the meaning of each formula. Eq. (9) forces k exemplars are identified for both topics T_A and T_B respectively, and Eq. (12) guarantees that m entity pairs are selected as the final result. The restriction given by Eq. (10) means each entity must choose only one exemplar. Eq. (11) enforces that if one entity e_j^X is voted by at least one other entity, then it must be an exemplar (i.e., $m_{jj}^X = 1$). The constraint given by (13) and (14) jointly guarantee that if an entity is selected in any comparable entity pair (i.e., $m_{ij}^T = 1$), then it must be an exemplar in its own topic (i.e., $m_{ii}^A = 1$ and $m_{jj}^B = 1$). Restricted by Eq. (15) and Eq. (16), each selected exemplar in the result is only allowed to appear once to avoid redundancy. $T'(M_X)$ depicts the overall typicality of selected exemplars in both topics T_A and T_B, and $G(e_i^X)$ denotes the representing subgroup for entity e_i^X (if e_i^X is not chosen as an exemplar, its representing subgroup will be null). $C'(M_T)$ depicts the overall comparability of generated entity pairs. In view of the fact that there are $2k$ numbers (each number is in [0,1]) in the typicality part $T'(M_A) + T'(M_B)$, and m numbers (each number is in [0,1]) in the comparability part $C'(M_T)$, we add the coefficients m and $2k$ in the objective function to avoid suffering from skewness problem. Finally, the parameter λ is used to balance the weight of the two parts. Our proposed ILP formulation is quite flexible, and guarantees to achieve the optimal solution.

6. Experiments

6.1 Datasets

We perform experiments on Wikipedia categories of different types including *location*, *person* and *organization*. In particular, the location categories compared are Japanese prefectures and US states (denoted by A_1, A_2, respectively). The person categories compared are Japanese prime ministers and US presidents(denoted by B_1, B_2, respectively). The organization categories compared are several top Japanese universities and US universities (denoted by C_1, C_2, respectively). The basic statistics of our datasets are shown in Tab. 1

Table 1 Summary of datasets.

Dataset	Wikipedia Category	# Entities
A1	Japanese Prefectures	46
A2	US States	50
B1	Japanese PMs	12
B2	US Presidents	45
C1	Japanese Universities	28
C2	US Universities	20

6.2 Baselines

To compare with our proposed ILP model for selecting exemplars and identifying comparable entity pairs, we first test our *hypothesis 1* and *hypothesis 2* on entity typicality discussed in Sec. 4. As clustering analysis aims to group objects into subsets and find the centroid of each group, we then compare our model with three widely-used clustering methods: K-Means clustering, DBSCAN clustering and aforementioned affinity propagation. Finally, as *stimulus similarity* suggests, typical exemplars are those who are similar to the other members of its category and dissimilar to members of the contrast categories. We adopt the mutually-reinforced random walk model to embody the idea of *stimulus similarity* and compare its performance with our method.

We brief discussed prepared five baselines below.

(1) ***Simple Typicality (ST)*** denotes the typicality measure for entities without implementing *hypothesis 1* and *hypothesis 2* in Sec. 4. Then we estimate an entity e given its input belonging set S by its likelihood of appearing in S $L(e|S)$, which is computed by Eq. (3).

(2) ***K-Means Clustering (K-Means)*** is a popular method used for cluster detection. It partitions all entities into clusters in which each entity belongs to the cluster that has the nearest mean. In this study we regard the entities closest to the centroids of their belonging clusters as exemplars, and we test whether such exemplars as typical.

(3) ***DBSCAN Clustering (DBSCAN)*** is a density-based clustering method which uses the concept of "core points" to represent points with high density. We test whether such "core points" are "typical exemplars". Note that we achieve the number of required clusters by adjusting the parameters *MinPts* and *Eps*, which are the minimal number of points within a radius from a core point, and the length of such radius, respectively.

(4) ***Affinity Propagation (AP)*** views the clustering as identifying a subset of representative exemplars. However, it does not guarantee the optimal solution. We test whether such selected exemplars are optimally typical exemplars. Note that we adjust the parameter *Preference* to satisfy the pre-defined number of clusters, where points with larger values of preferences are more likely to be chosen as exemplars.

(5) ***Mutually-Reinforced Random Walk (MRRW)*** uses a two-layer graph to compute typicality of entities of two compared categories. Each entity is represented as a node and entities belonging to the same category are located in the same layer. Besides, each edge between two nodes within the same layer is weighted by their similarity, and between two nodes of different layers is weighted by their dissimilarities. By within- and between-layer propagation in the graph, the scores from different layers can be mutually reinforced so that entities that get high scores tend to similar to entities within the same category and dissimilar to entities of the contrast category. Such entities are selected as exemplars.

After exemplars are chosen by each baseline above, we construct the entity pairs from selected exemplars. The constructed pairs are guaranteed to have the maximal score of sum of comparability between each pair.

6.3 Experiment Results

We display the experiment results by all the tested methods as follows. Tab. 2 shows the generated comparable entity pairs from dataset A_1 and A_2. Tab. 3 presents the results over datasets B_1 and B_2. Tab. 4 describe the results over datasets C_1 and C_2. The number of latent subgroups is set to be 5 for all the categories.

7. Conclusions

Comparison and learning from exemplars are effective strategies for obtaining comprehensive contrastive knowledge in the daily life. In this work, we propose a novel system to automatically detecting typical comparable entity pairs from two sets of entities. We adopt a concise ILP model for max-

Table 2 Top-5 Comparables generated over datasets A_1 and A_2 using each tested method.

Method	Entity Pair
ILP	(California, Tokyo), (Connecticut, Kanagawa), (Hawaii, Okinawa), (Kentucky, Gumma), (South::Dakota, Tottori)
ST	(Connecticut, Osaka), (Hawaii, Okinawa), (Idaho, Gunma), (Iowa, Chime), (Tennessee, Chiba)
K-Means	(Connecticut, Kanagawa), (Alabama, Chiba), (Iowa, Hiroshima), (Colorado, Chime) (Virginia, Nara)
DBSCAN	(Alaska, Hokkaido), (California, Kanagawa), (Maine, Tottori), (Ohio, Gunma), (South::Dakota, Yamanashi)
AP	(Hawaii, Okinawa), (Alaska, Hokkaido), (New::York, Tokyo), (California, Kanagawa), (Maine, Tottori)
MRRW	(South::Dakota, Ibaraki), (Missouri, Shizuoka), (Iowa, Kumamoto), (Arkansas, Chiba), (Colorado, Niigata)

Table 3 Top-3 Comparables generated over datasets B_1 and B_2 using each tested method.

Method	Entity Pair
ILP	(Millard::Fillmore, Yukio::Hatoyama), (Grover::Cleveland, Yamagata::Aritomo), (Franklin::D.::Roosevelt, Fumimaro::Konoe)
ST	(Grover::Cleveland, Yukio::Hatoyama), (James::K.::Polk, Shinzo::Abe), (Rutherford::B.::Hayes, Taro::Aso)
K-Means	(Herbert::Hoover, Fumimaro::Konoe), (Grover::Cleveland, Yamagata::Aritomo), (Gerald::Ford, Shinzo::Abe)
DBSCAN	(George::W.::Bush, Junichiro::Koizumi), (George::H.::W.::Bush, Yasuhiro::Nakasone), (Franklin::D.::Roosevelt, Fumimaro::Konoe)
AP	George::W.::Bush, Junichiro::Koizumi), (George::H.::W.::Bush, Yasuhiro::Nakasone), (Franklin::D.::Roosevelt, Fumimaro::Konoe)
MRRW	(Grover::Cleveland, Yukio::Hatoyama), (Theodore::Roosevelt, Shinzo::Abe), (Rutherford::B.::Hayes, Taro::Aso)

Table 4 Top-3 Comparables generated over datasets C_1 and C_2 using each tested method.

Method	Entity Pair
ILP	(Harvard Univ., Sophia Univ.), (Columbia Univ., International Christian Univ.), (Stanford Univ., Keio Univ.)
ST	(Stanford Univ., Univ. of Tokyo), (Columbia Univ., Kyoto Univ.), (Harvard Univ., Osaka Univ.)
K-Means	(MIT, Univ. of Tokyo), (Stanford Univ., International Christian Univ.), (Yale Univ., Waseda Univ.)
DBSCAN	(MIT, Univ. of Tokyo), (California Institute of Technology, Nagoya Univ.), (Stanford Univ., Kyoto Univ.)
AP	(MIT, Univ. of Tokyo), (California Institute of Technology, Nagoya Univ.), (Stanford Univ., Kyoto Univ.)
MRRW	(Stanford Univ., Tokyo Univ.), (Columbia Univ., Kyoto Univ.), (Harvard Univ., Keio Univ.)

imizing the overall representativeness and comparability of selected entity pairs. The experiment results demonstrate the effectiveness of our model compared to several strong baselines.

In the future, we plan to test our model on more heterogeneous datasets where context of entities are more difficult to compare for such scenarios are more general. We will also try to modify our model for query-sensitive comparative summarization tasks due to the good flexibility of our proposed ILP framework.

8. Acknowledgments

This research and development work was supported by the MIC/SCOPE #171507010.

References

[1] Alpa Jain and Patrick Pantel. 2009. Identifying comparable entities on the web. In Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 16611664.

[2] Ziheng Jiang, Lei Ji, Jianwen Zhang, Jun Yan, Ping Guo, and Ning Liu. 2013. Learning open-domain comparable entity graphs from user search queries. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2339 2344.

[3] Jindal, Nitin, and Bing Liu. "Identifying comparative sentences in text documents." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[4] Jindal, Nitin, and Bing Liu. "Mining comparative sentences and relations." AAAI. Vol. 22. 2006.

[5] Huang, Xiaojiang, Xiaojun Wan, and Jianguo Xiao. "Learning to find comparable entities on the web." Web Information Systems Engineering-WISE 2012 (2012): 16-29.

[6] Li, Shasha, et al. "Comparable entity mining from comparative questions." IEEE transactions on knowledge and data engineering 25.7 (2013): 1498-1509.

[7] Liu, Jiahui, Earl Wagner, and Larry Birnbaum. "Compare&contrast: using the web to discover comparable cases for news stories." Proceedings of the 16th international conference on World Wide Web. ACM, 2007

[8] Etzioni, Oren, et al. "Web-scale information extraction in knowitall:(preliminary results)." Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[9] McCallum, Andrew, and David Jensen. "A note on the unification of information extraction and data mining using conditional-probability, relational models." (2003).

[10] Sarawagi, Sunita, and William W. Cohen. "Semi-markov conditional random fields for information extraction." Advances in neural information processing systems. 2005.

[11] Feldman, Ronen, et al. "Extracting product comparisons from discussion boards." Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007.

[12] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[13] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[14] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.

[15] Xiao, Jianxiong, et al. "Joint affinity propagation for multiple view segmentation." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[16] Yu, Hai-Tao, et al. "A Concise Integer Linear Programming Formulation for Implicit Search Result Diversification." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.

[17] Garnett, Mathew J., et al. "Systematic identification of genomic markers of drug sensitivity in cancer cells." Nature 483.7391 (2012): 570-575.

[18] Zhang, Yating, et al. "Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time." ACL (1). 2015.

[19] Hua, Ming, et al. "Efficiently answering top-k typicality queries on large databases." Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007.

[20] Hua, Ming, et al. "Top-k typicality queries and efficient query answering methods on large databases." The VLDB Journal—The International Journal on Very Large Data Bases 18.3 (2009): 809-835.

[21] Au Yeung, Ching-man, and Ho-Fung Leung. "A formal model of ontology for handling fuzzy membership and typicality of instances." The computer journal 53.3 (2008): 316-341.

[22] Scott, David W., and Stephan R. Sain. "9-Multidimensional Density Estimation." Handbook of statistics 24 (2005): 229-261.

[23] Breiman, Leo, William Meisel, and Edward Purcell. "Variable kernel estimates of multivariate densities." Technometrics 19.2 (1977): 135-144.

[24] Rosch, Eleanor. "Cognitive representations of semantic categories." Journal of experimental psychology: General 104.3 (1975): 192.

[25] Brooks, Lee R. "Nonanalytic concept formation and memory for instances." (1978).

[26] Cohen, Benjamin, and Gregory L. Murphy. "Models of concepts." Cognitive science 8.1 (1984): 27-58.

[27] Barsalou, Lawrence W. "The instability of graded structure: Implications for the nature of concepts." Concepts and conceptual development: Ecological and intellectual factors in categorization 10139 (1987).

[28] Nosofsky, Robert M. "Similarity, frequency, and category representations." Journal of Experimental Psychology: Learning, Memory, and Cognition 14.1 (1988): 54-65.

[29] Tamma, Valentina, and Trevor Bench-Capon. "An ontology model to facilitate knowledge-sharing in multi-agent systems." The Knowledge Engineering Review 17.1 (2002): 41-60.

[30] Dubois, Didier, Henri Prade, and Jean- Paul Rossazza. "Vagueness, typicality, and uncertainty in class hierarchies." International Journal of Intelligent Systems 6.2 (1991): 167-183.