Ordering Concepts Based on Common Attribute Intensity

Tatsuya Iwanari
The University of Tokyo
nari@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
the University of Tokyo
ynaga@tkl.iis.u-tokyo.ac.jp

Nobuhiro Kaji*
Yahoo Japan Corporation
nkaji@yahoo-corp.jp

Toshiharu Nishina
Rakuten Inc.
toshiharu.nishina@rakuten.com

Masashi Toyoda
Institute of Industrial Science,
the University of Tokyo
toyoda@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
National Institute of Informatics
Institute of Industrial Science,
the University of Tokyo
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract

This paper presents a novel task of ordering given concepts (*e.g.*, *London*, *Paris*, and *Rome*) on the basis of common attribute intensity expressed by a given adjective (*e.g.*, *safe*) and proposes statistical ordering methods that integrate heterogeneous evidence extracted from text on concept ordering. This study is aimed at deriving collective wisdom on concept ordering from social media text. Solving this task is not only interesting from a sociological perspective but also beneficial in the practical sense for those who want to order unfamiliar entities in terms of subjective attributes that are hard to quantify in order to make correct decisions. Experiments on real-world concepts revealed a strong correlation between orderings obtained by our methods and gold-standard orderings.

1 Introduction

We make decisions every day by ordering two or more concepts on the basis of common knowledge or common sense to which we are privy. For example, imagine a situation in which we buy fruit juice. If we want something sweet to drink, we choose apple juice rather than lemon juice because we know that apples are generally sweeter than lemons.

The main objective of this study is to examine whether we can derive such common views on concept ordering from written text in social media, which reflect the perception of the crowd. Answering this question is not only interesting from a sociological perspective but also practically beneficial to those who want to order unfamiliar entities in terms of subjective attributes that are hard to quantify (*e.g.*, ordering smartphones in terms of *user-friendliness* or ordering tourist areas in terms of *fashionableness*) in order to make a correct decision. To come up with convincing ordering, we are presently forced to spend a substantial amount of time reading massive amounts of text to sum up people's perceptions or call for votes from domain experts.

Considering these situations in mind, we propose a novel task of ordering nominal concepts in accordance with the intensity of their common attributes as specified by adjectives. A set of nominal concepts (*e.g.*, {*elephant*, *whale*, *dog*, *mouse*}) is provided in the proposed task, along with an adjective (*e.g.*, *large*[1]) that represents an attribute shared by all members of the set. Given these two inputs, our goal is to output an ordered list of the items. The expected output in this example is *whale* $\succ$ *elephant* $\succ$ *dog* $\succ$ *mouse*, where the whale is the largest, the elephant is the second largest, and so forth.

An issue to be addressed in performing this task is how to define a gold-standard, or goal ordering. Since a concept could refer to various instances that have different attribute intensities, and some attributes are subjective, it is inherently difficult to define an absolutely correct ordering agreed on by all. We therefore asked multiple volunteers to order a given set of concepts and then use the ordering that achieved the best average Spearman [1904]'s rank correlation coefficient, ρ, against the human orderings as a gold-standard ordering. The resulting ordering can be seen as a representative ordering that sums up the general human perception, and is thereby meaningful as goal ordering in our task.

We present a method of ordering concepts to solve this task on the basis of textual evidence obtainable from massive amounts of social media text. The issues we address are twofold: (1) what kind of textual evidence to exploit

*This work was done in part while the author was at the Institute of Industrial Science, the University of Tokyo.

[1]We provide an antonym of a given adjective (*e.g.*, *small*) if any exists to reduce the ambiguity of the adjective.

and (2) how to integrate multiple kinds of evidence to obtain an appropriate ordering. We exploit heterogeneous textual evidence to address the first issue that indicates a possible ordering and then integrate this evidence to obtain an appropriate ordered list of the items. The types of evidence we used include noun-adjective (i.e., concept-attribute) co-occurrences, noun-adjective dependencies, similes, and comparative expressions on nouns. The first three indicate the absolute strength of the attribute intensity, while the last captures the relative strength among the attributes of concepts.

We explore two approaches to integrating the heterogeneous evidence to address the second issue. The first uses a pairwise learning-to-rank framework, specifically, a ranking support vector machine (SVM) [Joachims, 2002], while the second directly estimates Spearman's ρ for each candidate order to output the ordering with the highest estimated Spearman's ρ as the most likely ordering using a support vector regression (SVR) [Drucker *et al.*, 1997].

We performed experiments to evaluate our methods in terms of correlation between the system-generated and the gold-standard orderings for real-world concepts obtained from blog text. The results demonstrated that both our methods outperformed a co-occurrence-based baseline that was inspired by Turney [2002]'s work.

2 Related Work

To the best of our knowledge, there have been no attempts to order concepts on the basis of the intensity of their attributes. Related tasks are discussed in this section.

Question answering systems extract answers to factual questions (*e.g.*, '*What is the average temperature in Tokyo?*') from text [Prager, 2007]. Similarly, some researchers have attempted to extract objects' attributes and their values from the Web [Chen *et al.*, 2000; Yoshida *et al.*, 2003; Auer and Lehmann, 2007; Wu and Weld, 2007; Yoshinaga and Torisawa, 2007; Takamura and Tsujii, 2015]. These studies can partly help us to perform our task, particularly when we order concepts in terms of the intensity of objective and numerical attributes (*e.g.*, *largeness*, *heaviness*, and *expensiveness*).

Aspect-based sentiment analysis mines reviews or other texts for opinions on entities (*e.g.*, products or movies) [Pang and Lee, 2008]. Some of these studies have handled statements comparing multiple items (*e.g.*, '*car x is two feet longer than car y*' [Jindal and Liu, 2006]). Kurashima *et al.* [2008] proposed aggregating such statements to rank products in accordance with their popularity. This sort of information is also used with our method but is integrated with other evidence to obtain orderings for concepts that are not directly compared in texts. This strategy distinguishes our method from those proposed for aspect-based sentiment analysis.

In contrast to these studies, our task is more general in that it handles not only objective attributes (with numerical intensity, *e.g.*, *size* [Takamura and Tsujii, 2015]) but also subjective attributes. Further, it handles not only entities (with specific values for attributes) but also concepts (with a range of values for attributes).

There have been a range of studies on aggregating pairwise comparisons (partial orderings) to a single consensus ordering [Bradley and Terry, 1952; Volkovs and Zemel, 2012; Niu *et al.*, 2013; Chen *et al.*, 2013; Raman and Joachims, 2014]. These studies assume pairwise comparisons that are prepared (*e.g.*, search aggregation in meta-search or student evaluations via peer grading) or available from crowdsourcing, while we do not assume them in our task setting to increase the applicability of the method.

Făgărășan *et al.* [2015] proposed a method of inducing feature norms [McRae *et al.*, 2005] for a concept from text, which we use to perceive the concept. The features include adjectives (*e.g.*, 'is_sweet' for sugar) but exclude ordering expressions (*e.g.*, 'is_larger_than_a_pencil' for dog) so their task is complementary to our task, which helps us to suggest possible attributes for given concepts.

3 Method

We resort to massive amounts of social media text to collect textual evidence that validates our perception on concept ordering (Section 3.1) by assuming that our common views on concept ordering implicitly or explicitly affect the text we write. We then integrate that evidence to obtain a complete order of the concepts in the framework of supervised learning (Section 3.2).

3.1 Evidence on Concept Ordering

We exploit four types of evidence in this study to enable effective concept ordering. The first three implicitly suggest the absolute intensity of the attribute that the concept has. The fourth directly indicates a order of concepts, which captures the relative attribute intensity.

All pieces of evidence are represented by contexts where one or more concepts appear with a given adjective. The evidence is language-independent and we can tailor corresponding clue expressions for our target language, although our dataset is in Japanese. The example sentences that follow are provided in English to increase the applicability of our method and help reader understanding.

Noun-adjective co-occurrence If the intensity of the attribute of a concept is strong, we are likely to mention the attribute intensity along with the concept, which results in more sentences that include both the concept (noun) and the attribute (adjective).

- *Look how large that* [*elephant*] *is!*

Noun-adjective dependency A dependency relation between a nominative concept (noun) and an attribute (adjective) directly indicates the attribute intensity.

- [*Elephants*] *are so big*.

We used J.DepP [Yoshinaga and Kitsuregawa, 2009; 2010; 2014],[2] a state-of-the-art dependency parser, to extract such dependency relations. This evidence is less frequent but provides stronger evidence than co-occurrences, since co-occurrences do not always indicate the intensity of the attribute (*e.g.*, '*Ants are so small that* [*elephants*] *cannot harm them*').

[2] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

Simile If the intensity of the attribute of a concept is salient, we refer to the concept in a simile.

- *He is as brave as a lion*.

We use a couple of lexico-syntactic patterns tailored to detect similes, 'ADJ as NP' and 'ADJ like NP' ('*marude* NP *no you ni* ADJ' in Japanese).

Comparative This evidence directly indicates an order of a subset of a concept set. For example, the sentence below indicates *elephant* $\succ$ *dog*.

- *Elephants are larger than dogs*.

We use dependency relations for comparative adjectives with clues studied in the literature [Jindal and Liu, 2006] (a post particle '*yori*' in Japanese).

3.2 Two Approaches to Ordering Concepts

We explore two approaches to integrating the different types of heterogeneous evidence described in Section 3.1 to order concepts. The first uses a ranking support vector machine (SVM) [Joachims, 2002] to obtain an order and then induces concept-wise features from the evidence. The second uses a support vector regression (SVR) [Drucker *et al.*, 1997], which learns a function that directly maps a given order to Spearman's ρ against gold-standard ordering and induces order-wise features for candidate ordering.

Ordering Concepts with Ranking SVM

In ranking SVM, we represent each item (here, concept) with concept-wise features and perform pairwise training that generates $\mathcal{O}(n^2)$ training examples (feature vectors) from all the pairs of n concepts in the gold-standard order. The ranking SVM then solves an optimization problem that involves minimizing the number of incorrect partial orderings. Testing of the ranking SVM involves the dot product between feature and weight vectors for the input n concepts, and hence only requires $\mathcal{O}(n)$ time.

We encode the first three types of evidence described in Section 3.1 as real-valued features so that they directly indicate the attribute intensity of the concept. Generalizing Turney [2002]'s work that used the point-wise mutual information (PMI) of a pair of a word and 'excellent' (and 'poor') to compute the semantic orientation of the word, we measure the polarity of words in terms of attributes other than '*goodness*.' We compute feature value, $\phi(\boldsymbol{x})_{cooc}$, for the noun-adjective co-occurrence of a noun-adjective pair, $\boldsymbol{x} = (noun, adj)$, as:

$$\begin{aligned}\phi(\boldsymbol{x})_{cooc} &= SO^{adj}_{cooc}(noun) \\ &= \text{PMI}(noun, adj) - \text{PMI}(noun, \overline{adj}) \\ &= \log \frac{p(noun, adj)p(\overline{adj})}{p(noun, \overline{adj})p(adj)} \qquad (1)\end{aligned}$$

Here, $\overline{adj}$ refers to the antonym adjective or the adjective with negation. The feature values for dependency and simile are analogously computed, using co-occurrence counts based on dependency and simile.

We then encode the comparative expressions as two real-valued features so that they directly indicate the attribute intensity of the concept. For concept c and the adjective (here, *large*), we examine whether a comparative expression, 'c *is larger than* c',' can be found for any other concept $c'(\neq c)$ in the given set of concepts. We encode the number of c' found in the expression divided by the number of items in the given concept set to obtain one real-valued feature, which indicates the positive intensity of the attribute. We analogously compute another real-valued feature by looking at a comparative expression, 'c' *is larger than* c,' to express the negative intensity of the attribute.

Ordering Concepts with SVR

In SVR, we represent each possible ordering by order-wise features and then directly map it to a real-valued measure. The SVR can incorporate the evaluation measure into the ordering process and directly optimize an ordering by outputting the ordering with the maximum estimated measure. We use Spearman's ρ against the gold-standard order as a target variable for regression.

An issue to be addressed here is how candidate orderings for training and testing SVR are generated. There are factorial orders of candidate orderings, $n!$, for a given set of n concepts. As we do not need to use all the possible orderings in training, we adopted a Monte Carlo method that randomly samples only a tractable number of orderings from all possible permutations of the given concept set. We extract the same number of candidate orderings from each concept set to obtain balanced training data, and that number is set to $n_{min}!$, where n_{min} is the smallest number of members of all the concept sets in the training data. There were up to eight concepts for ordering in the experiments that follow, considering the annotation cost to order concepts and a typical application scenario where users order concepts that are chosen (or narrowed down) from candidates *a priori*. We thus consider all possible candidate orderings in testing to understand the maximum performance of this approach. We can greedily add one remaining concept to the (ordered) list by choosing the best insertion position in accordance with the SVR value of the resulting ordering in testing with a set of a large number of concepts, starting from an empty list. This requires $\mathcal{O}(n^2)$ time.

We induce real-valued features from the first three types of evidence described in Section 3.1 for the candidate ordering. We count the number[3] of ordered pairs (c, c') for which the $SO^{adj}_{cooc}(c)$ is larger than $SO^{adj}_{cooc}(c')$, of all the ordered pairs in the candidate ordering. This is used to examine the extent to which pairwise orderings in the candidate ordering conform to the ordering specified by SO^{adj}_{cooc}. The feature values for dependency and simile are analogously computed using co-occurrence counts based on dependency and simile.

We then induce two real-valued features from comparative expressions by counting the number[3] of ordered pairs (c, c') in which c's intensity is larger (or smaller) than c''s, in all the ordered pairs in the candidate ordering for the given adjective. Here, we assume the attribute intensity of a concept is given by the corresponding feature value in the ranking SVM. A larger feature value means that the candidate ordering satisfies (or dissatisfies) more partial orders found in the text.

[3]These numbers are normalized by dividing them by the number of all the ordered pairs in the candidate ordering.

Category	Adjective	Gold-standard ordering
flower	beautiful	sakura, rose, lily, lavender, platycodon, sunflower, camellia, daisy
jewel	elegant	sapphire, emerald, pearl, ruby, amethyst, opal, tourmaline, turquoise
alcohol	delicious	beer, wine, champagne, shōchū, chūhai, highball, tequila, makgeolli
sports	entertaining	football, table tennis, basketball, sumo, tennis, volleyball, baseball, professional wrestling
mammal	clever	dog, whale, cat, elephant, mouse, horse
mammal	large	whale, elephant, horse, dog, cat, mouse
conveyance	comfortable	Shinkansen, taxi, airplane, bicycle, bus, train
conveyance	fast	airplane, Shinkansen, train, taxi, bus, bicycle
food	yummy	steak, ramen, pasta, curry, pizza, fried rice, hamburger
instrument	soothing	flute, cello, clarinet, organ, trumpet, guitar, harmonica, drum
programming	easy	Ruby, Python, Perl, Java, JavaScript, Lisp, Scala, Haskell
programming	slow	Ruby, Perl, Python, JavaScript, Lisp, Haskell, Scala, Java
animal	lovely	squirrel, rabbit, dog, penguin, panda, horse, lizard, lion
vegetable	tasty	spinach, onion, pumpkin, eggplant, broccoli, napa cabbage, cucumber, sprout
fruit	sweet	melon, peach, apple, cherry, strawberry, tangerine, apricot
fruit	small	cherry, strawberry, apricot, tangerine, peach, apple, melon
appliance	useful	smartphone, PC, digital camera, car navigation system, printer, camera, speaker
flesh	preferable	chicken, beef, pork, lamb, brawn, venison, horseflesh
bird	cute	penguin, owl, quail, sparrow, swan, chicken, pheasant, eagle
weather	unpleasant	yellow sand, rain, thunder, gale, mist, snow, frost, fine
country	safe	UK, Thailand, Spain, India, Russia
country	warm	India, Thailand, Spain, UK, Russia
temple	famous	Kinkaku-ji, Ginkaku-ji, Hōryū-ji, Yakushi-ji, Zenkō-ji,Chūson-ji, Tō-ji, Zōjō-ji
temple	old	Hōryū-ji, Zenkō-ji, Yakushi-ji, Tō-ji, Chūson-ji, Zōjō-ji, Kinkaku-ji, Ginkaku-ji
cartoon	amusing	Gundam, Dragon Ball, One Piece, Vagabond, Kochikame, Gatchaman, Yatterman, Oishinbo
manufacturer	famous	Sony, Panasonic, Toshiba, NEC, Hitachi, Fujitsu, Canon, Seiko Epson
MLB team	famous	NYY, SEA, BOS, LAD, NYM, CWS, BAL, CLE
fast-food chain	tasty	MOS Burger, Freshness Burger, KFC, Mister Donut, Burger King, McDonald's
automaker	healthy	Toyota, Honda, Yamaha, Mazda, Daihatsu
corner store	useful	7-Eleven, Lawson, FamilyMart, Seicomart, Ministop
corner store	numerous	7-Eleven, Lawson, FamilyMart, Ministop, Seicomart
browser	friendly	Chrome, Firefox, Safari, Opera, Sleipnir
city	safe	London, Berlin, Paris, Hong Kong, Chicago, Rome, Moscow
coffee shop	likable	Starbucks, Saint Marc, Tully's, Pronto, Doutor, Excelsior, Ginza Renoir
town	fashionable	Aoyama, Shibuya, Shinjuku, Shinagawa, Nakano, Ikebukuro, Ueno, Asakusa

Table 1: Evaluation dataset.

4 Experiments

We performed experiments to evaluate our methods with open-domain datasets in terms of correlation between the system-generated and gold-standard orderings. We used LIBLINEAR [Fan *et al.*, 2008][4] as implementations of ranking SVM and SVR (with all hyper-parameters respectively tuned by cross-validation on training data).

4.1 Data

We used around 260 million Japanese blog articles, which we have crawled since 2006, to build a dataset for evaluation and obtain the statistics for our methods. The blog articles consist of around two billion sentences written by more than a million users.

We built an evaluation dataset for this task from the blog articles to include real-world concepts that are often absent in handcrafted ontologies such as WordNet. We first applied word clustering [Brown *et al.*, 1992] to one tenth of the blog articles in 2009 and obtained word clusters. We then looked into each cluster to manually collect nominal concepts in the same semantic category. Next, we associated each set of concepts with one or two adjectives that represented common attributes by examining the average PMI between the adjective and each concept, ultimately obtaining 35 pairs of a set of concepts and adjective (Table 1). There were 7.0 concepts per set on average and 28 unique sets of concepts (seven sets of concepts are associated with two adjectives). The resulting dataset included general to specific concepts (or instances) in various open-domain categories, and objective to subjective attributes for ordering that varied from one category to another.

We then asked seven volunteers (three graduate students, three researchers including the second author, and one system engineer) to provide an ordering for each pair of concepts and attribute. We regarded an ordering, in all permutations of concepts, that maximized the average of Spearman [1904]'s rank correlation coefficient, ρ,[5] against the seven human orderings, to be a gold-standard ordering. The gold-standard ordering we obtained is listed in Table 1.

[4] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[5] Spearman's ρ measures the strength of the correlation between two ordered lists. It ranges from -1 to 1. The negative value indicates an inverse correlation.

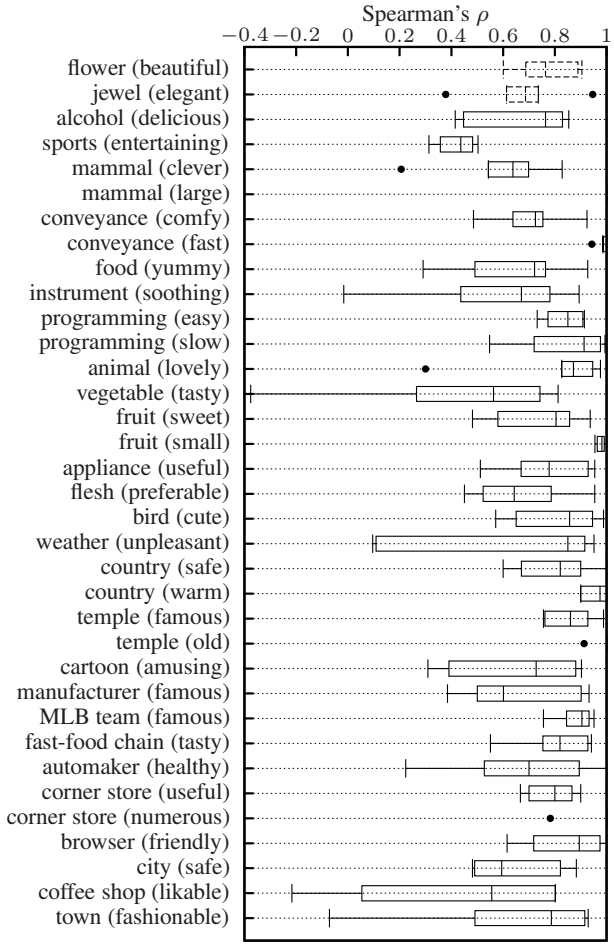

Figure 1: Correlation between the human orderings and the gold-standard ordering.

We computed Spearman's ρ between an ordering created by each of the seven subjects and the gold-standard ordering for each concept set to find the agreement between each human ordering and the gold-standard ordering. The results are summarized in a box-and-whisker diagram (Figure 1). We should state that the average Spearman's ρ between each human ordering and the gold-standard ordering was 0.75, which indicates a strong correlation.

We can see that the average Spearman's ρ between each human ordering and the gold-standard ordering is larger than 0.4 for all the pairs of concepts and adjective, larger than 0.6 for 30 pairs except for '*sports (entertaining)*,' '*mammal (clever)*,' '*instrument (soothing)*,' '*vegetable (tasty)*,' and '*coffee shop (likable)*,' and even larger than 0.8 for 15 pairs that included subjective ones such as '*programming (easy)*,' '*bird (cute)*,' and '*temple (famous)*,' which confirms that the perception of the relative intensities of common attributes is largely shared by human subjects. This demonstrates that the gold-standard ordering is appropriate as a common view, shared by human volunteers, on ordering for each pair.

Category (adjective)	HUMAN	BASE-LINE	SVM	SVR
flower (beautiful)	0.767	0.190	0.357	0.167
jewel (elegant)	0.682	0.286	0.524	0.548
alcohol (delicious)	0.663	0.000	0.762	0.810
sports (entertaining)	0.422	0.238	0.381	−0.095
mammal (clever)	0.598	0.371	0.143	0.029
mammal (large)	1.000	0.943	0.771	0.886
conveyance (comfy)	0.712	0.600	0.486	0.257
conveyance (fast)	0.986	0.257	0.543	0.771
food (yummy)	0.639	0.429	0.607	0.464
instrument (soothing)	0.583	0.405	0.310	0.238
programming (easy)	0.845	0.167	0.619	0.643
programming (slow)	0.840	0.214	0.381	0.238
animal (lovely)	0.806	0.738	0.548	0.595
vegetable (tasty)	0.462	0.786	0.524	0.476
fruit (sweet)	0.729	0.964	0.607	0.607
fruit (small)	0.979	0.143	0.607	0.536
appliance (useful)	0.772	0.036	0.393	0.500
flesh (preferable)	0.662	−0.286	0.143	−0.286
bird (cute)	0.819	0.905	0.929	0.929
weather (unpleasant)	0.664	0.714	0.524	0.143
country (safe)	0.804	−0.600	−0.200	0.000
country (warm)	0.961	0.900	0.700	0.700
temple (famous)	0.861	−0.214	0.524	0.619
temple (old)	0.988	0.095	0.595	0.667
cartoon (amusing)	0.648	−0.286	0.429	0.476
manufacturer (famous)	0.659	0.833	0.619	0.286
MLB team (famous)	0.885	0.690	0.905	0.976
fast-food chain (tasty)	0.807	0.886	0.886	0.543
automaker (healthy)	0.665	−0.600	−0.700	−0.900
corner store (useful)	0.791	−0.300	0.100	0.100
corner store (numerous)	0.969	0.500	0.500	0.500
browser (friendly)	0.856	−0.600	−0.600	−0.600
city (safe)	0.655	−0.357	0.357	0.250
coffee shop (likable)	0.405	0.143	0.786	0.464
town (fashionable)	0.673	0.405	0.381	0.262
average	0.750	0.274	0.441	0.366

Table 2: Results on ordering concepts: Spearman's ρ against gold-standard ordering.

4.2 Results

We conducted leave-one-out cross-validation using the evaluation dataset described in Section 4.1. The appropriateness of the system-generated orderings was then measured by computing Spearman's ρ between the system-generated and gold-standard orderings. The experimental results are listed in Table 2. Here, HUMAN refers to the average Spearman's ρ between each human ordering and the gold-standard ordering, while BASELINE refers to a method that scores each concept by using the PMIs of noun-adjective co-occurrences for the given adjective and its antonym (SO^{adj}_{cooc}, i.e., Eq. 1), as was done in Turney [2002]'s work. SVM and SVR refers to our methods based on ranking SVM and SVR (Section 3.2), respectively. The results revealed that our methods overwhelmed the baseline, which indicated the effectiveness of integrating heterogeneous evidence in a supervised manner. Ranking SVM achieved the best average performance and indicated a positive correlation for all pairs other than '*country (safe)*,' '*automaker (healthy)*,' and '*browser (friendly)*.'

Table 3 shows an ablation test that evaluates the impact of individual pieces of evidence in ranking SVM. All the ev-

Method	Spearman's ρ
SVM (all)	0.441
−co-occurrence	0.391
−dependency	0.407
−simile	0.292
−comparative	0.424

Table 3: Ablation test for ranking SVM.

idence contributed to improving the correlation against the gold-standard orderings. The simile most significantly affects the performance, followed by noun-adjective co-occurrences and dependencies. This conforms to our expectation since similes are useful for finding concepts with strongest (or weakest) attribute intensity, which greatly contributed to the improvement of Spearman's ρ against the gold-standard orderings. The comparative expressions are rarely observed in the text, so that had the least impact.

We manually investigated the system-generated orderings to identify which evidence had contributed to the orderings. Table 4 lists examples of system-generated orderings. Noun-adjective dependencies between concept (noun) and *famous* in the ordering of '*MLB team (famous)*' raised the ordering of *SEA* and *NYY*, so the resulting orderings of our systems correlated with the gold-standard ordering. Similes in the ordering of '*fruit (small)*' greatly contributed to improving the ordering of *cherry* (to top-1), which again increased the performance of our systems. Comparative expressions in the ordering of '*programming (easy)*' contributed to raising the ordering of *Java*, which again increased the performance of our systems. These changes suggest that the use of heterogeneous evidence increases the chances of reaching ordering that has a higher correlation with the gold-standard ordering, which justifies our approach to integrating the pieces of evidence in a supervised manner.

The low correlation in '*country (safe)*' and '*automaker (healthy)*' was caused by their status changes over our time-series text. As for '*browser (friendly)*,' similes and comparative expressions were not observed, and this data sparseness reduced the correlation. There were also some difficult cases in gathering reliable evidence. For example, we sometimes mention unusual and surprising things such as "*The mice around here are large*," which have a negative impact on ordering concepts.

	GOLD	BASELINE	SVM	SVR
MLB team (famous)				
1	NYY	SEA	SEA	NYY
2	SEA	BOS	NYY	SEA
3	BOS	NYM	BOS	BOS
4	LAD	NYY	NYM	NYM
5	NYM	CWS	CWS	LAD
6	CWS	BAL	LAD	CWS
7	BAL	LAD	BAL	BAL
8	CLE	CLE	CLE	CLE
alcohol (delicious)				
1	beer	makgeolli	beer	wine
2	wine	beer	wine	beer
3	champagne	shōchū	shōchū	champagne
4	shōchū	highball	champagne	shōchū
5	chūhai	champagne	makgeolli	makgeolli
6	highball	wine	highball	chūhai
7	tequila	tequila	tequila	tequila
8	makgeolli	chūhai	chūhai	highball
fruit (small)				
1	cherry	apricot	cherry	cherry
2	strawberry	tangerine	apricot	apricot
3	apricot	peach	tangerine	apple
4	tangerine	cherry	apple	tangerine
5	peach	apple	peach	peach
6	apple	melon	strawberry	strawberry
7	melon	strawberry	melon	melon
programming (easy)				
1	Ruby	JavaScript	Ruby	Ruby
2	Python	Scala	Perl	Java
3	Perl	Perl	Java	Perl
4	Java	Python	Scala	JavaScript
5	JavaScript	Ruby	JavaScript	Scala
6	Lisp	Java	Python	Python
7	Scala	Haskell	Haskell	Haskell
8	Haskell	Lisp	Lisp	Lisp

Table 4: Examples of system-generated orderings.

5 Conclusion

We have initiated a novel task of ordering concepts on the basis of the intensity of their common attributes and proposed methods of ordering a given set of concepts by aggregating heterogeneous pieces of evidence obtained from massive amounts of social media text. As the experimental results obtained from real-world concepts revealed a strong correlation between the system-generated and gold-standard orderings, we could induce common views on ordering concepts from texts written by the crowd. These results are not only interesting from the sociological perspective but are also beneficial in practice since they would help us make decisions from among alternative concepts in daily life.

This research has involved various directions for productive future research. First, we plan to support more specific and diverse attributes that are described by phrases other than adjectives (*e.g.*, *easy to cook*) to order concepts (*e.g.*, recipes). Since supporting more specific attributes makes the data sparseness problem more serious, we are going to incorporate correlations between attribute intensities (*e.g.*, heavier concepts are likely to be larger, or more expensive concepts are likely to have better quality). Next, we intend to apply our method to text written in different periods (in 2014 vs. in 2015), in different geographical areas (Japan vs. UK), or by different demographics (people in their 30's vs. 50's or males vs. females), to observe changes in common views over time or place, or by different demographics.

We will release the evaluation dataset (Table 1) with human orderings and the experimental codes for the academic and industrial communities at http://www.tkl.iis.u-tokyo.ac.jp/~nari/ijcai-16/ to facilitate the reproducibility of our results and their use in various application contexts.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 25280111, 16K16109 and 16H02905. The authors wish to thank the volunteers for their hard work. The authors also thank the anonymous reviewers for their valuable comments.

References

[Auer and Lehmann, 2007] Sören Auer and Jens Lehmann. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. In *Proceedings of the 4th European Conference on The Semantic Web (ESWC)*, pages 503–517, 2007.

[Bradley and Terry, 1952] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, December 1952.

[Brown *et al.*, 1992] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December 1992.

[Chen *et al.*, 2000] Hsin-Hsi Chen, Shih-Chung Tsai, and Jin-He Tsai. Mining tables from large scale HTML texts. In *Proceedings of the 18th conference on Computational linguistics (COLING)*, pages 166–172, 2000.

[Chen *et al.*, 2013] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 193–202, 2013.

[Drucker *et al.*, 1997] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir N. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9, NIPS 1996*, pages 155–161. MIT Press, 1997.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Făgărășan *et al.*, 2015] Luana Făgărășan, Eva Maria Vecchi, and Stephen Clark. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57, 2015.

[Jindal and Liu, 2006] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *Proceedings of the 21st national conference on Artificial intelligence (AAAI)*, pages 1331–1336, 2006.

[Joachims, 2002] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

[Kurashima *et al.*, 2008] Takeshi Kurashima, Katsuji Bessho, Hiroyuki Toda, Toshio Uchiyama, and Ryoji Kataoka. Ranking entities using comparative relations. In *Proceedings of the 19th Conference on Database and Expert Systems Applications (DEXA)*, pages 124–133, 2008.

[McRae *et al.*, 2005] Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005.

[Niu *et al.*, 2013] Shuzi Niu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. Stochastic rank aggregation. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 478–487, 2013.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc., 2008.

[Prager, 2007] John Prager. *Open-Domain Question Answering*. Now Publishers Inc., 2007.

[Raman and Joachims, 2014] Karthik Raman and Thorsten Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1037–1046, 2014.

[Spearman, 1904] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, January 1904.

[Takamura and Tsujii, 2015] Hiroya Takamura and Jun'ichi Tsujii. Estimating numerical attributes by bringing together fragmentary clues. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1305–1310, 2015.

[Turney, 2002] Peter Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.

[Volkovs and Zemel, 2012] Maksims N. Volkovs and Richard S. Zemel. A flexible generative model for preference aggregation. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*, pages 479–488, 2012.

[Wu and Weld, 2007] Fei Wu and Daniel S. Weld. Autonomously semantifying Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pages 41–50, 2007.

[Yoshida *et al.*, 2003] Minoru Yoshida, Kentaro Torisawa, and Jun'ichi Tsujii. Extracting attributes and their values from Web pages. In *Web Document Analysis: Challenges and Opportunities*, Series in Machine Perception and Artificial Intelligence, pages 179–200. World Scientific Publishing, Inc., 2003.

[Yoshinaga and Kitsuregawa, 2009] Naoki Yoshinaga and Masaru Kitsuregawa. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1542–1551, 2009.

[Yoshinaga and Kitsuregawa, 2010] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel slicing: Scalable online training with conjunctive features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1245–1253, 2010.

[Yoshinaga and Kitsuregawa, 2014] Naoki Yoshinaga and Masaru Kitsuregawa. A self-adaptive classifier for efficient text-stream processing. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1091–1102, 2014.

[Yoshinaga and Torisawa, 2007] Naoki Yoshinaga and Kentaro Torisawa. Attribute-value acquisition from semi-structured texts. In *Proceedings of the Workshop of OntoLex07 - From Text to Knowledge: The Lexicon/Ontology Interface held at the sixth International Semantic Web Conference*, pages 55–66, 2007.