Regular Paper

DBSJ Journal Vol. 16, No. 2
March 2018

Inducing Writers' Values on Concept Ordering from Microblog

Tatsuya IWANARI[♡]
Naoki YOSHINAGA[◇]
Masashi TOYODA[♠]
Masaru KITSUREGAWA[♣]

This article proposes a robust method of inducing microblog writers' values on concept ordering in a specific domain (*e.g.*, *genders*, *residential areas* and *time series*) from their writings in the domain. The values on concept ordering are represented by sets of ordered concepts (*e.g.*, *London*, *Berlin*, and *Rome*) in accordance with a common attribute intensity expressed by an adjective (*e.g.*, *entertaining*). Existing methods infer social-media users' values by aggregating various pieces of evidence for the given concepts and adjective from their writings, but suffer from a data sparseness problem when a target domain becomes more specific since it is more difficult to gather a sufficient amount of evidence from less data. We therefore introduce two techniques to solve the data sparseness problem: 1) exploiting adjectives whose intensity correlates with that of the target adjective (*e.g.*, *heavy* for *large*) and 2) referring to concept orderings in more general domains where more text is available than the target domain. We evaluate our method on real-world concept orderings with various domains on our 5-year microblog (Twitter) archive.

1. Introduction

We make decisions every day by ordering two or more concepts on the basis of common knowledge or common sense that we have. 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. On the other hand, when we want to investigate unfamiliar things or concepts (*e.g.*, Gypsophila), we typically endeavor to understand the concept by comparing or ordering it with similar and familiar concepts (*e.g.*, Rose and Carnation) from various perspectives (*e.g.*, beautiful, cheap) and specific standpoints (*e.g.*, for women, in spring). At present, people are forced to spend a substantial amount of time wading through massive amounts of text to get an overview of others' opinions, or spend a lot of money to call for votes from experts or crowd workers in order to derive a convincing ordering.

Motivated from these situations, Nishina et al. [15] initiated a task of ordering concepts based on common attribute intensity expressed by an adjective (§ 2), and Iwanari et al. presented a system that derives concept orderings [10] by aggregating various pieces of evidence such as co-occurrence of a concept and adjective from social media text [9] (§ 3). The system collects microblog posts written by specific writers and at a certain time of interest (say, *domain*) to induce concept orderings of the target domain, which reflect their values on the target concepts. It is not only practically beneficial for understanding concepts from others' ordering-based values to make correct decisions but also interesting from a sociological perspective for inversely understanding common views shared by a certain demographic and/or from a certain period of time. As the target domain becomes more specific, however, it becomes more difficult to gather enough amount of evidence due to the data sparseness problem, which prevents the system from making convincing orderings.

To solve the data sparseness problem, we propose a robust method of ordering concepts that gathers more evidence by (1) exploiting adjectives whose intensity correlates with that of the target adjective and (2) referring to concept orderings in more general domains (where more text is available) than the target domain in the supervised framework [10] (§ 4). Addressing the data sparseness problem, this study opens a way to acquiring values in more specific domains, or ultimately, individual values.

We validate the effectiveness of our method in terms of the correlation between the system-generated and gold-standard orderings for real-world concepts obtained from social media text (§ 5). Experimental results on our 5-year Twitter archive confirmed that our method obtained more convincing concept orderings in specific domains than the baseline [9].

2. Task Settings

This section describes input, output, and gold standard of our concept ordering task. We exploit microblog posts in a specific domain to induce the common values shared by the writers (users) in that domain. The domains of individual users are identified in advance (§ 5.1.2).

Input A set of nominal concepts is provided along with an adjective that represents an attribute shared by all members of the set. We provide an antonym of the target adjective if any exists to reduce the ambiguity of the adjective. We refer to a pair of concepts and adjective (and its antonym) as a query. In addition to a query, our method accepts one of the pre-identified domains (*e.g.*, women, living in Kanto region).

Output Given these inputs, our goal is to output an ordered list of given concepts on the basis of attribute intensity. For example, when a set of concept {*elephant*, *whale*, *dog*, *mouse*} and an adjective *large* (along with the antonym *small*) are given, the expected output is *whale* ≻ *elephant* ≻ *dog* ≻ *mouse*, where *whale* is the largest, *elephant* is the second largest, and so forth. The output ordering is required to reflect the common values of writers in the specified domain.

♡ Non Member Recruit Holdings Co,Ltd.
nari@tkl.iis.u-tokyo.ac.jp

◇ Member Institute of Industrial Science, the University of Tokyo
ynaga@iis.u-tokyo.ac.jp

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

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

Gold Standard We ask multiple crowd workers to order concepts from various viewpoints (adjectives) and to provide their domain information (*e.g.*, age, gender, prefecture they live in, SNS they use) (§ 5.1.1). We then generate the gold-standard orderings for a domain that maximize the average Spearman's rank correlation coefficient, ρ, against the orderings of crowd workers in the domain. The resulting orderings can be considered as common values shared in the domain.

3. Related Work

The concept ordering is a relatively new task initiated by Nishina et al. [15]. In this section, we first discuss tasks related to the concept-ordering task, and then introduce existing approaches to the concept ordering.

Question answering systems extract answers to factual questions (*e.g.*, '*What is the average temperature in Tokyo?*') from text [18] and some researchers have attempted to extract attributes and their values from the Web [4, 1, 24, 28, 21]. These studies can partly help us to perform our task, particularly when we order concepts in terms of the intensity of an 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) [17]. Some of these studies have handled statements comparing multiple items (*e.g.*, '*car x is two feet longer than car y*') [11]. Kurashima *et al.* [13] 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, the concept ordering task is more general in that it handles not only objective attributes (with numerical intensity, *e.g.*, *size* [21]) but also subjective attributes. Furthermore, it handles not only entities (with specific values for attributes) but also concepts (with a range of values for attributes).

There has been a range of studies on aggregating pairwise comparisons (partial orderings) to a single consensus ordering [2, 23, 16, 5, 19]. 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.

Nishina et al. [15] initiated the concept ordering task that we tackled in this article and proposed a method that orders concepts on the basis of the point-wise mutual information (PMI) of noun-adjective dependencies, inspired by Turney's work [22]. Iwanari et al. [10] proposed methods that order concepts by gathering various pieces of evidence from social media text and integrating them with a supervised learning. The method outperformed Nishina's method and the authors confirmed that it is possible to obtain common views of whole social media user from the text people have written. Iwanari et al. further developed a system that helps interactively understand the values in different domains by retrieving posts to gather evidence in the target domain [9]. However, they did not address how to solve the data sparseness problem which occurs when a user wants to know values of more specific domains in which a smaller amount of text, thus a smaller amount of evidence, is available.

Our study addresses the above data sparseness problem by exploiting adjectives correlating with the target adjective and statistics obtained from domains that are more general than the target domain. Lee et al.'s recent work [14] (which appeared after the draft version of this article first appeared) addresses the data sparseness problem by exploiting synonymous adjectives in a way similar to our approach. They obtained promising results for ordering concepts with English datasets, which confirms the effectiveness of using similar adjectives to solve the data sparseness problem.

4. Proposed Method

This section describes our method of concept ordering. We adopt Iwanari's supervised framework of concept ordering [10], and introduce two smoothing techniques to gather supplemental pieces of evidence on concept ordering to address the data sparseness problem. The first technique exploits evidence on adjectives whose intensity correlates with that of the target adjective (*e.g.*, *heavy* for *large*), while the second one refers to evidence obtained from domains that are more general than the target domain.

In what follows, we first briefly explain Iwanari's method of concept ordering (§ 4.1). We then explain the two smoothing techniques (§ 4.2, § 4.3).

4.1 Ordering Concepts Based on Common Attribute Intensity

Iwanari et al. [10] resorted to massive amounts of social media text to collect textual evidences that represent writers' perception on concept ordering, and then obtained a convincing ordering by integrating these evidences in ranking SVM [12] and support vector regression (SVR) [6]. They exploited four types of evidences to capture the common view on concepts from social media text: (1) co-occurrences of a concept and an adjective (*e.g.*, How large that whale is!), (2) dependencies from a concept to an adjective (*e.g.*, A whale is so big.), (3) similes (*e.g.*, He is brave as a lion.), and (4) comparative expressions (*e.g.*, Whales are larger than cats.). The first three types of evidence implicitly suggest attribute intensity and can be understood as capturing the absolute intensity of the attribute that the concept has. The fourth directly captures the relative attribute intensity, which directly indicates the order of a subset of a concept set. These four types of evidence are encoded as real-valued features by using the point-wise mutual information (PMI) of the pairs of a concept and adjective for each piece of evidence. As an extension to this method, Iwanari et al. [9] developed a system to infer values in a specific domain by gathering posts from specific segments of microblog writers (*e.g.*, genders, regions) and/or using posts in different time periods.

In this study, we gather microblog posts from specific domains in the same way, adopt these four types of evidence and order concepts with ranking SVM, since they reported that ranking SVM worked better than SVR.

4.2 Evidence from Correlating Adjectives

We exploit adjectives whose intensity correlates with that of given adjectives and collect the basic four types of evidence for given concepts and the correlating adjectives. To expand a given adjective, we use a method that scores candidate ad-

Table 1: Query set (79 queries: 41 unique categories / 48 unique adjectives).

Category	Concepts	Adjectives
bird	fowl, swan, penguin, owl, sparrow	large, cute
vegetable	spinach, cucumber, sprout, onion, Chinese cabbage, eggplant, pumpkin	healthy, delicious
fruit	strawberry, orange, apple, melon, cherry, persimmon, grape	sweet, large
mammal	dog, bear, whale, mouse, lion	clever, large
jewelry	pearl, sapphire, opal, garnet, turquoise	elegant, rare
instrument	cello, flute, violin, clarinet, harp	graceful, pleasant
flower	cherry, sunflower, bellflower, lily of the valley, dandelion	beautiful, likable
cafe	Doutor, Saint Marc, Tully's, Komeda, Ginza Renoir	delicious, expensive
manufacturer	Sony, Panasonic, Toshiba, Fujitsu, Canon, Seiko Epson, Hitachi	well, new
country	Thailand, India, the United Kingdom, Russia, Spain, the United States, China	wealthy, vast, warm
automaker	Toyota, Honda, Yamaha, Mazda, Daihatsu	well, famous
alcohol	high ball, beer, chuhai, whiskey, sake	delicious, expensive
food	hamburger, noodles, fried rice, curry, pizza	likable, fatty
appliance	printer, washer, car navigation system, cameras, air conditioner	more expensive, noisy
weather	rain, snow, thunder, fog, strong wind, frost, clear sky	likable, rare
flesh	beef, pork, chicken, lamb, horsemeat	likable, more expensive
temple	Ginkakuji, Zenkoji, Yakushiji, Chusonji, Zojoji, Toji	famous, magnificent
sport	table tennis, basketball, tennis, volleyball, football, baseball, sumo	major, good at
conveyance	airplane, Shinkansen, train, taxi, bus	comfortable, fast, safe
actress	Ki Kitano, Tomochika, Rinka, Yumiko Shaku, Yuka, Akina Minami, Kazue Fukiishi	cute, interesting
cake	short cake, cheese cake, roll cake, chocolate cake, chiffon cake	sweet, likable
baked goods	macaroon, scone, bagel, muffin, sponge cake	delicious, fashionable
drink	powdered tea, black tea, cocoa, green tea, orange juice	delicious
specialty	sanuki udon, okonomiyaki, curry rice with pork cutlet, beef tongue, kushikatsu	likable
food	sanuki udon, okonomiyaki, curry rice with pork cutlet, beef tongue, kushikatsu	more expensive
city	Tokyo, Osaka, Fukuoka, Nagoya, Kobe, Okinawa, Sapporo	warm, distant
cuisine	Chinese, Thai, Spanish, Korean, Indian	healthy, spicy
profession	police officer, doctor, scientist, astronaut, composer	capable, harsh
movie	Alice in Wonderland, Beauty and the Beast, My Neighbor Totoro, Nausicaä of the Valley of the Wind, Star Wars	interesting, new
subject	mathematics, English, physical education, Japanese, world history	indispensable, easy
leisure	reading, fishing, jogging, surfing, BBQ, driving	pleasant, meaningful, easy
media	Youtube, Instagram, Facebook, Twitter, Niconico	interesting, convenient
anime	Dragon Ball, JoJo, Pretty Cure, Sailor Moon, Eva, GTO	interesting
politician	Shinzo Abe, Taro Aso, Yukio Hatoyama, Junichiro Koizumi, Kakuei Tanaka	young, likable
foreign company	Apple, Google, Yahoo, Samsung, Microsoft	well, essential
celebrity	Edison, Kenji Miyazawa, Prince Shotoku, Ryoma Sakamoto, Newton	great, likable
entertainer	Takeshi Beat, Sanma Akashiya, Tamori, George Tokoro, Shinsuke Shimada	interesting, young
tourist site	Lake Biwa, Izumo Taisha, Tsutenkaku, Osaka Castle, the Imperial Palace	precious
era	Edo period, Yayoi period, Heian period, Nara period, Kamakura period	new, long
characteristic	hairstyle, clothes, looks, kindness, speech	important
male athlete	Ichiro Suzuki, Kei Nishikori, Yuzuru Hanyu, Darvish, Uchimura Kohei	young, wonderful

jectives by calculating the PMI of dependencies from a *candidate* to the *target*. In the scoring process, we also consider the polarity of adjectives by considering not only the given adjectives but also their antonyms, and handling negations by extending Turney's work [22] (Equation (1)), which is also used to calculate the feature values of the evidence (§ 4.1).

$$\begin{aligned} &SO^{adj}_{dep}(candidate) \\ &\quad = \text{PMI}(\underline{ad\,jective\ or\ not\ antonym}, candidate) \\ &\qquad - \text{PMI}(\underline{antonym\ or\ not\ ad\,jective}, candidate) \end{aligned} \tag{1}$$

Note that we use PMI of neither adjectives co-occurrences nor dependencies from the *target* to a *candidate* since co-occurrences were noisy and the dependencies occasionally imply the opposite cause-and-effect relationship between the two adjectives. For instance, "this orange is sweet and tasty" can imply sweet things are tasty. We thereby count up sweet as a *candidate* when the *target* is tasty but we do not count up tasty as a *candidate* when the *target* is sweet in this case.

We regard expanded adjectives which have the best (or worst) K scores for the target adjective as its correlating adjectives. We then accumulate the evidence counts of K expanded adjectives to form single features. We ignore some noisy adjectives such as '*good*' and '*bad*' because they make a bad influence on ordering by occupying a majority of evidence counts. As the number of the basic evidence type is four, we now have another set of four types of evidence for the set of K extended adjectives. We have released the tool at `https://github.com/tiwanari/pmi-box`.

4.3 Evidence from General Domain

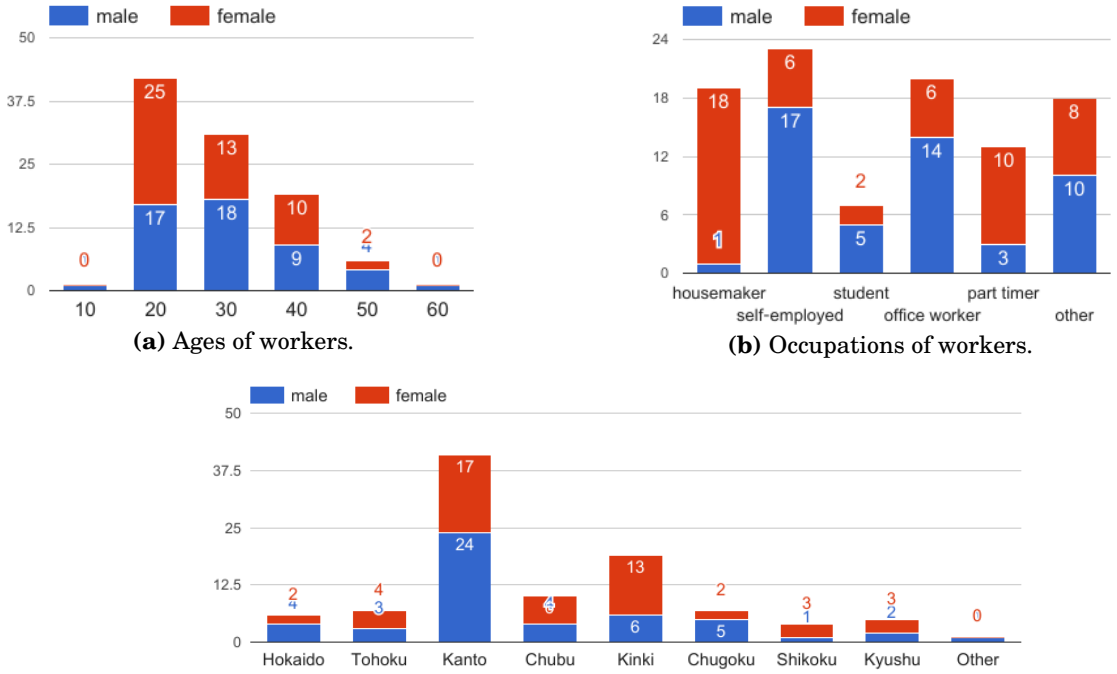
Assuming that you are a female Twitter user who lives in Tokyo, you must have some tastes in common with other female users, and with other users who live in Tokyo. We make use of this intuition by referring to orderings of more general domains than the target domain as a sort of a prior.

Having a target domain as an input along with a set of concepts and adjective, our method collects statistics not only from the target domain but also from more general domains to compute feature vectors per domain on the basis of the statistics. As the result, we have d more feature vectors, where d refers to the number of domains that are more general than the target domain. We then concatenate them $(\vec{v_1}, \ldots, \vec{v_d})$ with the feature vector of the target domain $(\vec{v}_{target})$ as shown in Equation (2) and use this extended vector $(\vec{v_{ex}})$ for training and testing.

$$\vec{v_{ex}} = (\vec{v}_{target}, \vec{v_1}, \ldots \vec{v_d}) \tag{2}$$

5. Evaluation

In this study, we built an evaluation dataset for concept ordering by crowdsourcing (Table 1) and evaluated our method in terms of the correlation between the system-generated and gold-standard orderings. We used in-house Twitter



(a) Ages of workers.

(b) Occupations of workers.

(c) Regions of workers (Other is out of Japan).

Figure 1: The domain of crowd workers: the blue and red numbers show the number of male and female workers, respectively.

Table 2: Evaluation datasets and correlation between human orderings. (·) shows the number of workers in each domain.

	GENERAL			KANTO			KINKI		
	ALL (100)	FEMALE (50)	MALE (50)	ALL (41)	FEMALE (17)	MALE (24)	ALL (19)	FEMALE (13)	MALE (6)
Ave. ρ	0.588	0.595	0.602	0.604	0.634	0.614	0.608	0.611	0.696

archive (detailed in § 5.1.2) to collect pieces of evidence for ordering. We used LIBLINEAR [7][1] as an implementation of ranking SVM (with all hyper-parameters respectively tuned by cross-validation on training data). In the following sections, we tried to obtain ordering-based values of users in different genders and/or in different areas such as male Twitter users who live in KANTO region, Japan.

5.1 Data

5.1.1 Evaluation Datasets

We generated 79 queries with the same process of Iwanari et al. [9], which used Brown clustering [3], on our 2012's Japanese blogs (about 165 million sentences) and Twitter archive (about 3 billion tweets) to include a wide variety of concepts and adjectives: from concepts (*e.g.*, '*airplane*') to instances (*e.g.*, '*Ginkakuji*', a temple) and from objective adjectives (*e.g.*, '*fast*') to subjective ones (*e.g.*, '*likable*'). The list of all the queries is shown in Table 1.

After preparing the query set, we gathered 100 Japanese Twitter or blog writers by a crowdsourcing service[2] and asked them to answer (rank) each query to create gold-standard orderings for training and testing. The crowd workers had diverse demographics: gender (50 males and 50 females), age (from 20s to 60s), location (30 out of 47 prefectures in Japan) and occupation (students, homemakers, office workers, etc.). Figure 1 summarized their demographics information.

We generated gold-standard orderings for each domain by choosing an ordering from all the permutations of concepts, which maximized the average of Spearman's rank correlation coefficient ρ [20] against the orderings of the workers in the domain. The correlations of some domains are shown in Table 2. Here, ALL refers to the average Spearman's ρ between the gold-standard ordering and all crowd worker orderings, while FEMALE and MALE refer to the average ρ among female and male crowd workers, respectively. In addition to them, we calculated the average ρs for data with KANTO and KINKI tags that were gathered only from microblog writers living in KANTO region (Ibaraki, Tochigi, Gumma, Saitama, Chiba, Tokyo, Kanagawa prefectures) and KINKI region (Mie, Shiga, Kyoto, Osaka, Hyogo, Nara, Wakayama prefectures), respectively. The gold-standard orderings have enough strong correlations against human orderings (around 0.60) and the gold-standard orderings of more specific domains have higher average correlations; namely, crowd workers in a more specific domain agree more with their gold-standard orderings in the domain. Looking into the correlations, we can see the differences in ordering-based values between domains. For example, for a query '*alcohol (delicious)*,' women have much stronger correlation than men have. We have released the whole list of correlations on our website to promote the repli-

[1] https://www.csie.ntu.edu.tw/~cjlin/liblinear/
[2] https://crowdworks.jp/

Table 3: The number of users living in each region.

Region	# users
HOKKAIDO	6,905
TOHOKU	7,696
KANTO	117,023
CHUBU	18,430
KINKI	31,895
CHUGOKU	5,498
SHIKOKU	2,034
KYUSHU	11,893

cability of our results at `http://www.tkl.iis.u-tokyo.ac.jp/~nari/deim-17/`.

5.1.2 Twitter Datasets

We have crawled Twitter posts for more than five years by using Twitter API since March 11, 2011. We started crawling timelines from 26 famous Japanese users, and then repeatedly expanded the set of users by following retweets and mentions appeared in the timelines while tracking their timelines. Our archive has more than 2 million users and 25 billion tweets.

Next, we briefly analyzed gender and location of the Twitter users from their posts and profiles in order to annotate posts with their domains. For gender, we adopted a simple heuristic that determines the gender according to the number of clue expressions (in their posts) indicating either gender; the clue expressions include first-person pronouns and sentence-ending particles that are specific to each gender [8]. For location, we exploited the user profiles to annotate the location (living prefecture) of users. We extracted common locations specified in their profiles by sorting the locations according to their frequency. We then manually assigned the common locations to an appropriate prefecture. The gender classifier detected 345 thousand males and 311 thousand females (Japanese users), and the region classifier detected 201 thousand Japanese users. Table 3 shows the detail of identified regions and the number of users. Here, note that the distribution of Twitter users' region data (Table 3) is similar to that of crowd workers' region data (Figure 1c). This is because they were randomly sampled and reflect the population distribution of Japan, and therefore they are suitable for evaluation.

We used 2012-2016 data from the archive to gather evidence because they contain whole year tweets and thus are free from time series biases which have been seen in [9]. In the evidence gathering process, we counted concept-adjective co-occurrences per tweet instead of per sentence and used J.DepP [25, 26, 27],[3] a state-of-the-art dependency parser, with mecab-ipadic-NEologd[4] to extract dependency relations.

5.1.3 Expanded Adjectives

We expanded the given adjectives of the evaluation data as explained in § 4.2. We used our 2012's Japanese blog archive which contains about 165 million sentences. The blog articles have more formal expressions compared to Twitter posts and we can thereby extract more reliable correlating adjectives. Table 4 lists obtained adjectives (translated into English) with the best and worst 3 SO^{adj}_{dep} values for each target adjective.

[3] `http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/`
[4] `https://github.com/neologd/mecab-ipadic-neologd`

Table 4: Expanded adjectives (the best and worst 3).

Adjective	Best 3	Worst 3
large	suitable, moderate, fine	simple, light, weak
cute	fine, bright, young	outright, desperate, true
healthy	delicious, easy, yummy	superficial, hot, dubious
delicious	tender, irresistible, fragrant	clumsy, poor, unstable
sweet	easy, sour, thick	easy, lucky, lovely
clever	easy, cute, kind	troublesome, stupid, free
elegant	simple, neat, friendly	endless, stupid, inferior
rare	awesome, funny, cute	similar, plenty, heaviness
graceful	beautiful, elegant, delicate	cheap, natural, large
pleasant	kind, cool, natural	terrible, creepy, unpleasant
beautiful	bright, wide, vivid	foolish, lowly, shallow
likable	cool, pretty, funny	fickle, dismal, cramped
expensive	distant, powerful, wide	optimistic, ample, cheesy
well	small, unknown, high	sore, long, heavy
new	early, well, early	dark, narrow, feminine
wealthy	free, convenient, peaceful	terrible, desire, awesome
vast	rich, incredible, beautiful	fierce, narrow, round
warm	simple, thick, gentle	fierce, incredible, hot
famous	delicious, large, yummy	wise, plain, young
expensive	precious, beautiful, heavy	abundant, easy, facile
fatty	salty, delicious, sweet	expensive, wonderful, inaptness
noisy	persistent, smelly, sore	gentle, sparse, peaceful
magnificent	big, awesome, wide	best, exaggerated, tidy
major	easy, simple, sweet	sober, famous, small
good at	tight, strong, great	hate, troublesome, dark
comfortable	safe, wide, convenient	terrible, creepy, unpleasant
fast	overwhelming, accurate, sharp	busy, various, equivalent
safe	fresh, healthy, strong	sweet, remarkable, unstable
interesting	mysterious, distinctive, thrilling	cramped, fatal, empty
fashionable	simple, beautiful, cute	insensitive, few, pleasant
distant	rugged, endless, close	soft, unlimited, standard
spicy	moderate, difficult, tough	ambiguous, thick, distant
capable	numerous, high, awesome	depressed, disturbing, weird
harsh	long, tough, miserable	white, serious, undecided
indispensable	cold, hot, difficult	meaningless, awesome, magnificent
easy	unnecessary, healthy, facile	tough, hard, professional
pleasant	tasty, great, bright	efficient, grabby, suitable
meaningful	fun, valuable, many	few, subtle, distinctive
easy	cheap, easy, convenient	heavy, cheap, strong
convenient	close, easy, possible	unreliable, rapid, uniform
young	pervy, fine, beautiful	dull, poisonous, narrow
essential	important, fatigue, important	alien, abundant, simple
great	big, incredible, wonderful	thankful, lovely, noble
precious	few, many, fun	doubtful, heavy, sorry
long	endless, steep, complex	fleeting, close, danger
important	amazing, cheap, important	desirable, frustrating, plump
wonderful	fun, young, many	historical, firm, sound

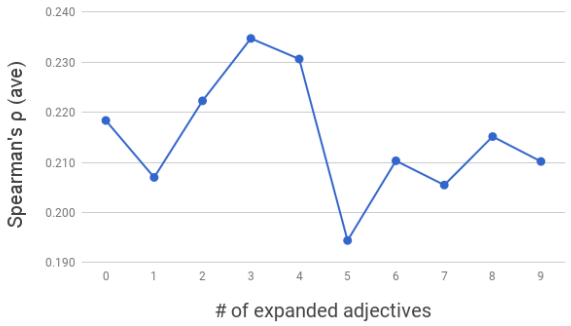


Figure 2: Different numbers of expanded adjectives.

5.2 Result

We conducted leave-one-out cross-validation using the evaluation dataset (§ 5.1.1) on our Twitter archive (§ 5.1.2). The appropriateness of the system-generated orderings was measured by computing Spearman's ρ between the system-generated and gold-standard orderings. We varied the number of expanded adjectives and confirmed using the best and worst 3 correlating adjectives achieved the best average ρ as shown in Figure 2. This illustrated using too many ex-

Table 5: Results on ordering concepts: Spearman's ρ is measured against gold-standard ordering.

(a) Results with general domains.

	ALL		FEMALE				MALE			
	BASE	+ADJS	BASE	+ADJS	+GEN	+BOTH	BASE	+ADJS	+GEN	+BOTH
Ave. ρ	0.237	**0.239**	0.196	0.197	**0.256**	0.197	0.185	**0.239**	0.202	0.222

(b) Results with specific (Kanto) domains.

	KANTO (ALL)				KANTO (FEMALE)				KANTO (MALE)			
	BASE	+ADJS	+GEN	+BOTH	BASE	+ADJS	+GEN	+BOTH	BASE	+ADJS	+GEN	+BOTH
Ave. ρ	0.262	**0.282**	0.261	0.262	0.290	**0.305**	0.290	0.276	0.211	**0.235**	0.227	0.198

(c) Results with more specific (Kinki) domains.

	KINKI (ALL)				KINKI (FEMALE)				KINKI (MALE)			
	BASE	+ADJS	+GEN	+BOTH	BASE	+ADJS	+GEN	+BOTH	BASE	+ADJS	+GEN	+BOTH
Ave. ρ	0.198	0.213	0.214	**0.227**	0.165	**0.188**	0.168	0.120	0.223	0.215	**0.240**	0.232

Table 6: General domains for the target domains.

Target domain	General domain(s)
ALL	-
FEMALE	ALL
MALE	ALL
KANTO (ALL)	ALL
KANTO (FEMALE)	ALL, FEMALE, KANTO (ALL)
KANTO (MALE)	ALL, MALE, KANTO (ALL)
KINKI (ALL)	ALL
KINKI (FEMALE)	ALL, FEMALE, KINKI (ALL)
KINKI (MALE)	ALL, MALE, KINKI (ALL)

panded adjectives are not helpful but harmful for making a convincing ordering since they could contain noisy and improper words.

We evaluated our method with nine domains which are the same domains explained in § 5.1.1 (ALL, FEMALE, MALE and these three with KANTO and KINKI tag). The experimental results are listed in Table 5. Here, BASE refers to Iwanari et al.'s method [9], which uses neither expanded adjectives nor general domain information, and +ADJS, +GEN and +BOTH refer to our method that extends the baseline with expanded adjectives (§ 4.2, with the best and worst 3 expanded adjectives), general domain information (§ 4.3) and both of them, respectively. For the +GEN method, we extended feature vectors by using orderings of more general domains than target domains and the list of these general domains is shown in Table 6. We used one general domain for FEMALE, MALE, KANTO (ALL) and KINKI (ALL), and three general domains for the others.

Because ALL does not have more general domains, +GEN (and +BOTH) cannot apply to the domain. The results showed that our two techniques worked better compared to the baseline. Here, note that we cannot compare the correlation of a domain with that of other domains because the gold-standard orderings are different and each domain has its own gold-standard. For Table 5a and 5b, +ADJS overwhelmed the baseline in all the cases and they had the best average ρ in the most of the domains. On the other hand, in Table 5c, which is more specific than others, +GEN worked better than the baseline and +ADJS (except KINKI (FEMALE)) and +BOTH underperformed the baseline in KINKI (FEMALE). Considering the number of the users and tweets in our Twitter archive, ALL and KANTO contain much larger amount of data than KINKI has and using expanded adjectives simply helped our method gather more evidence for these general domains. However, since KINKI does not have enough amount of data, the number of occurrences between expanded adjectives and the target concepts was still not enough large to compute reliable feature values, and the smoothing techniques did not improve the correlations very much. In the case, referring to general domains' information was a better way to obtain reliable orderings. In short, using expanded adjectives helped better for more general domains while general domain information worked better for more specific domains.

5.3 Case Studies and Error Analysis

We manually investigated the gold-standard and system-generated orderings in order to analyze errors and confirm the effectiveness of our method. Since the number of queries is too large (79 queries) to list all the results, we picked out some of them here and the full set of the gold-standard and system-generated orderings is available on our website.

We firstly analyzed errors of our methods. Referring to Table 4, we can see the expanded adjectives accidentally included some irrelevant adjectives and they would have been noisy for counting. For example, the method generated '*simple*' and '*thick*' as the best correlating adjectives of '*warm*' but they do not seem to correlate with and, to make matters worse, the method suggested '*hot*' as the third worst correlating adjective of '*warm*.' This definitely dropped the correlations of +ADJS for '*country (warm)*' in almost all the domains (KINKI (FEMALE) had no change), and the error decreased the correlation from 0.714 (BASE) to −0.607 (+ADJS) in the worst case (KANTO (MALE)). The problem can be solved by refining the expansion process. As for +GEN, it failed to solve '*electric appliance (noisy)*' and lowered the correlations in all the domains. For the query, the system generated the opposite ordering ($\rho = -1$) against the gold-standard ordering for ALL, which is the most general domain, and the error of general domains was propagated to specific domains. In this case, +BOTH gave good hints by referring to +ADJS rather than +GEN, and improved the correlations if +ADJS generated proper orderings.

We then show some examples of system-generated order-

Table 7: Examples of system-generated orderings. Spearman's ρs are measured against the gold-standard orderings.

	GOLD	BASE	+ADJS	+GEN	+BOTH
'*celebrity (great)*' - ALL					
ρ		0.700	0.900		
1	Edison	Edison	Edison		
2	Newton	Prince Shotoku	Newton		
3	Ryoma Sakamoto	Newton	Prince Shotoku		
4	Prince Shotoku	Ryoma Sakamoto	Ryoma Sakamoto		
5	Kenji Miyazawa	Kenji Miyazawa	Kenji Miyazawa		
'*baked goods (delicious)*' - KANTO (ALL)					
ρ		-0.500	0.700	0.500	0.500
1	sponge cake	macaroon	sponge cake	scone	scone
2	muffin	scone	scone	sponge cake	sponge cake
3	scone	sponge cake	bagel	bagel	bagel
4	bagel	bagel	muffin	muffin	muffin
5	macaroon	muffin	macaroon	macaroon	macaroon
'*flesh (expensive)*' - KINKI (MALE)					
ρ		0.000	0.500	0.700	1.000
1	beef	pork	horsemeat	beef	beef
2	horsemeat	lamb	pork	pork	horsemeat
3	lamb	horsemeat	beef	horsemeat	lamb
4	pork	beef	lamb	lamb	pork
5	chicken	chicken	chicken	chicken	chicken

ings (Table 7). Here, GOLD refers to the gold-standard ordering of a specified domain. The first example is '*celebrity (great)*' (ALL). For this query, both BASE and +ADJS achieved good correlations against the gold-standard ordering. +ADJS succeeded in gathering more pieces of evidence and made the better result by raising the order of Newton. Secondly, BASE generated a bad ordering for '*baked goods (delicious)*' (KANTO (ALL)) but the resulting orderings of our method were much more correlated with the gold-standard ordering by exploiting expanded adjectives and general domain information. Thirdly, for '*flesh (expensive)*' (KINKI (MALE)), BASE did not create a convincing ordering because a small amount of evidence was found in the domain. Our smoothing techniques outperformed the BASE by generating more convincing orderings, and +BOTH created the best ordering ($\rho = 1$). These examples confirmed that our method is effective to mitigate the data sparseness problem.

6. Conclusion

We proposed a robust method of ordering concepts by gathering evidence aggressively from social media text. The method helps to acquire the writers' ordering-based values, which are represented by sets of ordered concepts in accordance with a common attribute intensity expressed by an adjective, in specific domains where a small amount of ordering clues are available by exploiting 1) adjectives whose intensity correlates with that of target adjectives and 2) global information of more general domains than the target domain.

We evaluated our method with our 5-year Twitter archive and confirmed that our method overwhelmed the baseline and is helpful to improve the correlations between the system-generated and gold-standard orderings. Addressing the data sparseness problem, this study opened a way to inferring microblog writers' values in more specific domains. We confirmed that we need further improvements when we combine smoothing techniques through the evaluation.

We have released the evaluation dataset at `http://www.tkl.iis.u-tokyo.ac.jp/~nari/deim-17/`.

[Acknowledgments]

This work was partially supported by JSPS KAKENHI Grant Number 16K16109 and 16H02905.

[Bibliography]

[1] 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.

[2] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. volume 39, pages 324–345, December 1952.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 2008.

[8] Daisuke Ikeda, Tomoyuki Nanno, and Manabu Okumura. Gender classification of blog authors. In *Proceedings of the 12th Annual Meeting for Natural Language Processing*, pages 356–359, 2006. (in Japanese).

[9] Tatsuya Iwanari, Kohei Ohara, Naoki Yoshinaga, Nobuhiro Kaji, Masashi Toyoda, and Masaru Kitsuregawa. Kotonush: Understanding concepts based on

values behind social media. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING), System Demonstrations*, pages 292–296, 2016.

[10] Tatsuya Iwanari, Naoki Yoshinaga, Nobuhiro Kaji, Toshiharu Nishina, Masashi Toyoda, and Masaru Kitsuregawa. Ordering concepts based on common attribute intensity. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3747–3753, 2016.

[11] 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.

[12] 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.

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

[14] Kyungjae Lee, Hyunsouk Cho, and Seung won Hwang. Gradable adjective embedding for commonsense knowledge. In *Proceedings of the 21st Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 814–827, 2017.

[15] Toshiharu Nishina, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Ordering concepts using paired adjectives. In *IPSJ SIG Notes, NL-214*, number 8, pages 1–7, Nov 2013. (in Japanese).

[16] 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.

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

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

[19] 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.

[20] Charles Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.

[21] 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.

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

[23] 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.

[24] Fei Wu and Daniel S. Weld. Autonomously semantifying Wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM)*, pages 41–50, 2007.

[25] Naoki Yoshinaga and Masaru Kitsuregawa. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, 2009.

[26] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel Slicing: Scalable Online Training with Conjunctive Features. In *Proceedings of the 23rd conference on Computational linguistics*, pages 1245–1253. Tsinghua University Press, 2010.

[27] Naoki Yoshinaga and Masaru Kitsuregawa. A self-adaptive classifier for efficient text-stream processing. In *Proceedings of the 25th conference on Computational linguistics*, 2014.

[28] Naoki Yoshinaga and Kentaro Torisawa. Open-domain attribute-value acquisition from semi-structured texts. In *Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge: The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66, 2007.

Tatsuya IWANARI

He is currently an employee of Recruit Holdings Co,Ltd. He received his Bachelor's and Master's degree in information science and technology from the University of Tokyo, Japan in 2015 and 2017, respectively. His research interests include natural language processing, parallel distributed computing, and programming education for beginners.

Naoki YOSHINAGA

He received his BSc and MSc in information science and his PhD in Information Science and Technology from the University of Tokyo, Japan, in 2000, 2002, and 2005, respectively. He is an associate professor at the Institute of Industrial Science, the University of Tokyo, Japan. He was a JSPS research fellow (DC1, PD) from 2002 to 2008. He worked at the Institute of Industrial Science, the University of Tokyo, as a specially appointed associate professor from 2012 to 2016, and also worked at the National Institute of Information and Communications Technology (NICT), Japan, as a senior researcher from 2014 to 2016. His research interests include computational linguistics and natural language processing.

Masashi TOYODA

He received the PhD degree in computer science from the Tokyo Institute of Technology, Japan, in 1999. He is an associate professor of the Institute of Industrial Science, the University of Tokyo, Japan. He worked at the Institute of Industrial Science, the University of Tokyo, as a specially appointed associate professor from 2004 to 2006. His research interests include web mining, user interfaces, information visualization, internet of things, and visual programming.

Masaru KITSUREGAWA

He is currently the Director General of National Institute of Informatics (NII), and also a professor of the University of Tokyo. He received his Ph.D. degree in information engineering from the Univ. of Tokyo in 1983. He has been working in the area of high performance database system and systems for big data. He served as a president of Information Processing Society of Japan (IPSJ) and a science advisor for Ministry of Education, Culture, Sports, Science and Technology, Japan. He is a fellow of ACM, IEEE, IPSJ and IEICE.