DEIM Forum 2017 F6-1

Inducing Writers' Values on Concept Ordering from Microblog

Tatsuya IWANARI†, Naoki YOSHINAGA††, Masashi TOYODA††, and Masaru KITSUREGAWA††,†††

† Graduate School of Information Science and Technology, the University of Tokyo,
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8654 Japan
†† Institute of Industrial Science, the University of Tokyo,
4–6–1 Komaba, Meguro-ku, Tokyo, 153–8505 Japan
††† National Institute of Informatics, the University of Tokyo,
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–0003 Japan
E-mail: †{nari,ynaga,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract This paper proposes a robust method of ordering concepts to acquire ordering-based values of people in a specific domain (e.g., genders, residential areas and time series) or ultimately, values of individuals on the basis of what they write about on microblogs. The ordering-based values 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 work [5], [6] proposed statistical methods that infer writers' values from what they have written in social media. These methods suffer from the data sparseness problem since it becomes difficult to gather a sufficient amount of evidence to make a convincing ordering as the target domain becomes more specific. We therefore introduce two techniques to solve the data sparseness problem by: 1) exploiting adjectives (*e.g.*, *heavy*) whose intensity correlates with that of the target adjective (*e.g.*, *large*) and 2) refering 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 domains on our 5-year Twitter archive.

Key words Natural Language Processing, Social Media, Learning to Rank, Concept Ordering, Data Sparseness

1. Introduction

When we want to investigate unfamiliar entities or concepts (*e.g.*, iPhone SE) as consumers, or inversely, intend to supply new concepts as vendors, we typically endeavor to understand the values of a concept by comparing or ordering it with similar and familiar concepts (*e.g.*, Xperia X or Galaxy S7) from various perspectives (*e.g.*, *user-friendliness*) and for various domains (*e.g.*, which product men prefer). At present, people often spend a substantial amount of time wading through massive amounts of text to get an overview of others' perceptions, or spend a lot of money to call for votes from experts in order to come up with a convincing ordering.

We proposed a system that infers people's values on given concepts by gathering pieces of evidence from a large amount of social media text. The values can be inferred as concept orderings on the basis of common attribute intensity expressed by the target adjective [5]. The system collects posts from social media text written by specific users and at a certain time of interest (say, *domain*) to induce concept orderings specific to the target domain. The system is not only practically beneficial for understanding entities from others' ordering-based values to make correct decisions but also interesting from sociologically perspective for inversely understanding common views shared by a certain demographic and/or from a certain period of time.

However, as the target domain becomes more specific, the system suffers from the data sparseness problem and the problem prevents the method from making convincing orderings. To solve the problem, we propose a robust method of ordering concepts that uses two smoothing techniques that: (1) exploit adjectives whose intensity correlates with that of a given adjective and (2) refer to concept orderings in more general domains (where more text is available) than the target domain. Addressing the data sparseness problem, this paper 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 the gold-standard orders for real-world concepts obtained by Brown clustering [1] from Web text.

The remainder of this paper is organized as follows. Section 2. formally defines the task of our study. Section 3. introduces work related to our study. Section 4. presents our smoothing techniques that gather evidence aggressively. Section 5. evaluates our method on our Twitter dataset. Section 6. concludes this study and addresses future work.

2. Task Setting

We exploit social media text in specific domains to induce the common values shared by the users in the domains. The domains of users are identified in advance.

a) Input

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

b) Output

Given these inputs, our goal is to output an ordered list of the 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 of it *small*) are inputted, 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 in the specified domain by analyzing text written by the users in the domain.

c) Gold Standard

We ask multiple volunteers to order the given set of concepts from various viewpoints (adjectives) and to provide their domain information (*e.g.*, age, gender, prefecture they live in, SNS they use). We then generate the gold-standard orderings for a domain that maximize the average Spearman [10]'s rank correlation coefficient, ρ, against the orderings of volunteers in the domain. The resulting orderings can be considered as common values shared in a domain.

3. Related Work

To the best of our knowledge, there have been no attempts other than Nishina et al. [9]'s work (in Japanese) and our own previous work [5], [6] on ordering concepts on the basis of the intensity of their attributes.

Nishina et al. [9] initiated the task that we tackled in this paper and proposed a method that orders concepts on the basis of the point-wise mutual information (PMI) of noun-adjective dependencies inspired by Turney [12]'s work. We also use the information but we combine it with other evidence as features in the framework of supervised learning.

Iwanari et al. [6] 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 et al. [9]' method and they confirmed that it is possible to obtain common views of whole social media user from the text people have written.

Extending the methods, Iwanari et al. [5] developed a system to interactively understand the values in different domains by retrieving posts to gather evidence from the target domain. 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 is available.

This paper addresses the data sparseness problem by exploiting adjectives related to the target adjective and global information of the domains which are more general than the target domain.

4. Method

We aggressively gather the basic four types of evidence used in Iwanari et al. [6] with two smoothing techniques and integrate the evidence with ranking SVM [7]. The first smoothing technique exploits adjectives (*e.g.*, large) whose intensity correlates with that of the target adjective (*e.g.*, heavy) (Section 4. 2) and the other gathers global information of more general domains than the target domain (Section 4. 3). Firstly, We briefly explain Iwanari et al. [5], [6]'s work which is the basis of our study and then introduce our proposal.

4. 1 Ordering Method

Iwanari et al. [6] resorted to massive amounts of social media text to collect textual evidences that represent our perception on concept ordering and then obtained a convincing ordering by integrating these evidences in ranking SVM [7] and support vector regression (SVR) [2]. 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 evidences 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. [5] developed a system to infer values in a specific domain by gathering posts from specific segments of Twitter users (*e.g.*, genders, regions) and/or using posts in different time periods.

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

4. 2 Use of Correlating Adjectives

We exploit adjectives whose intensity correlates with that of given adjectives and collect the basic four types of evidence for them with the given concepts. To expand a given target adjective, we use a method that scores candidate adjectives by using the PMI of dependencies from a candidate to the target. In the scoring process, we also consider the polarity of adjectives by using not only the given adjectives but also their antonyms while handling negations by extending Turney [12]'s work (Equation 2), which is also used to calculate the feature values of the evidence described in Section 4. 1.

$$SO^{adj}_{dep}(candidate) \quad (1)$$

$$= \textsc{pmi}(\underline{adjective}\ or\ \underline{not\ antonym}, candidate) - \textsc{pmi}(\underline{antonym}\ or\ \underline{not\ adjective}, candidate) \quad (2)$$

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

We regard expanded adjectives which have best (or worst) K scores per target adjective with the above process as the 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' that occur far more times than the *target* because they occupy a majority of evidence counts and thus make a bad influence for ordering. 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 Exploiting General Domain Information

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

Receiving the 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 and computes feature vectors per domain on the basis of the statistics. As the result, we have d more feature vectors per concept, where d refers to the number of domains that are more general than the target domain. We then concatenate them ($\vec{v_1}$, ... , $\vec{v_d}$) with the feature vector of the target domain ($\vec{v}_{target}$) as shown in Equation 3 and use this extended vector ($\vec{v_{ex}}$) for training and testing.

$$\vec{v_{ex}} = (\vec{v}_{target},\ \vec{v_1},\ ...\ \vec{v_d}) \quad (3)$$

5. Evaluation

We evaluated our method with our Twitter archive in terms of the correlation between the system-generated and gold-standard orderings. We used LIBLINEAR [3] (`https://www.csie.ntu.edu.tw/~cjlin/liblinear/`) 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 in Iwanari et al. [5], which used Brown clustering [1] on our 2012's Japanese blog (about 165 million sentences) and Twitter archive (about 3 billion tweets) to include various kinds of concepts and adjectives. We have obtained a wide variety of concepts and adjectives as queries: 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 users by a crowdsourcing service (`https://crowdworks.jp/`) and asked them to answer (rank) each query to create gold-standard orderings for training and testing. The crowd workers had various 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 workers' demographics information.

We generated gold-standard orderings for each domain by choosing an ordering, in all permutations of concepts, that maximized the average of Spearman [10]'s rank correlation coefficient ρ 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 fe-

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

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

				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

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

male crowd workers and among male crowd workers, respectively. In addition to them, the average ρs were calculated for data with KANTO and KINKI tag that were gathered only from users 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, that is to say, the crowd workers in a more specific domain agree more with their gold-standard orderings in the domain. Therefore, looking into the correlations, we can see the differences between domains. For example, as for a query '*alcohol (delicious)*', women have much stronger correlation than men have. We will release the whole list of correlations on our website to promote the replicability of our result.

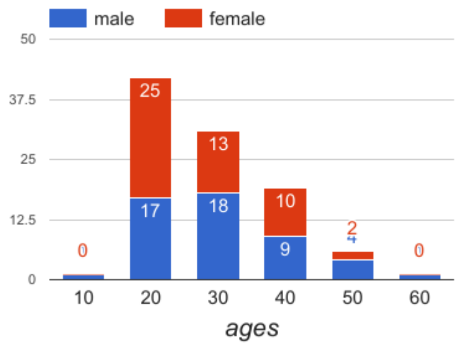
5.1.2 Twitter Datasets

We have crawled Twitter posts for more than five years by using Twitter API since March 11, 2011. We started crawling timelines from 30 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.

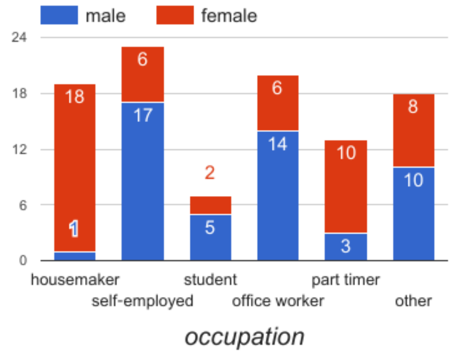
Region	# users
HOKKAIDO	6905
TOHOKU	7696
KANTO	117023
CHUBU	18430
KINKI	31895
CHUGOKU	5498
SHIKOKU	2034
KYUSHU	11893

Table 3: The summary of identified regions of users.

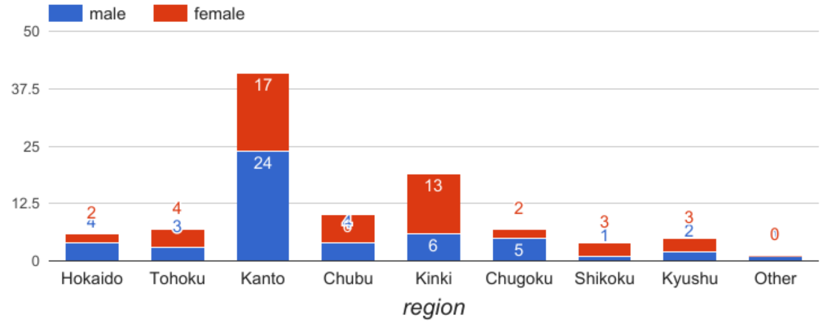
Next, we briefly analyzed gender and location of the Twitter users from their posts and profiles in order to annotate



(a) Ages of workers.



(b) Occupations of workers.



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

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

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 [4]. For location, we exploited the user profiles to annotate the location (living prefecture) of users. We extracted common locations specified by the users 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 tweet and thus are free from time series biases which have been seen in Iwanari et al. [5]. In the evidence gathering process, we counted concept-adjective co-occurrences per tweet not per sentence and used J.DepP [8], [13], [14] (`http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/`), a state-of-the-art dependency parser, along with mecab-ipadic-NEologd [11] to extract dependency relations.

5. 1. 3 Expanded Adjectives

We expanded the given adjectives of the evaluation data with our method explained in Section 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 and thus we can extract more reliable correlative adjectives. We listed candidate adjectives which have best 3 and worst 3 SO^{adj}_{dep} values per target adjective.

Table 4 shows the list of expanded adjectives for the evaluation datasets. We translated the Japanese adjectives into English.

5. 2 Results

We conducted leave-one-out cross-validation using the evaluation dataset (Section 5. 1. 1) on our Twitter archive (Section 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 evaluated our method with nine domains which are the same domains explained in Section 5. 1. 1 (ALL, FEMALE, MALE and these three with KANTO and KINKI tag). The ex-

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

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

perimental results are listed in Table 6. Here, BASE refers to the baseline which was used in Iwanari et al. [5] and +ADJS, +GEN and +BOTH refer to our method that extended the baseline with expanded adjectives (Section 4.2, best 3 and worst 3), general domain information (Section 4.3) and both of them, respectively.

As 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 5. We used one general domain for FEMALE, MALE, KANTO (ALL) and KINKI (ALL), and three general domains for the others.

The summary of the results is shown in Table 6. Because ALL does not have more general domains than it, +GEN (and +BOTH) cannot apply to the domain. The results showed that our two techniques worked well compared to the baseline. Here, note that we cannot compare the correlation

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)

Table 5: General domains for the target domains.

of a domain with that of other domains because the gold-standard orderings are different and each domain has their own gold-standard. As for Table 6a and 6b, +ADJS overwhelmed the baseline in all cases and they had the best average ρ in the most of the domains. On the other hand, in Table 6c, which is relatively 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 did not have enough amount of data, the number of occurrences between expanded adjectives and the target concepts was not enough large to compute reliable feature values and the smoothing technique did not improve the correlations very much. In such a case, referring to general domains' information was a better way to obtain reliable orderings.

5.3 Case Studies and Error Analysis

We manually investigated gold-standard and system-generated orderings in order to analyze errors and confirming the effectiveness of our method. Since the number of queries is too large (79 queries) to list all of the results, we picked out some of them here and the full set of the gold-standard and system-generated orderings will be 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 be noise 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 3rd worst correlating adjective of 'warm.' This surely 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) for the worst case (KANTO (MALE)). The problem can be solved

	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

(a) Results with general 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

(b) Results with specific (Kanto) 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

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

Table 6: Results on ordering concepts: Spearman's ρ against gold-standard ordering.

	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

Table 7: Examples of system-generated orderings. Spearman's ρs against the gold-standard orderings are shown for each ordering.

by refining the expanding process. As for +GEN, it failed to solve '*electric appliance (noisy)*' and lowered the correlations compared to BASE in all the domains. This can be explained that the error of general domains was propagated to specific domains. With the query, +GEN generated the opposite ordering ($\rho = -1$) against the gold-standard ordering for ALL, which is the most general domain and referred by all other 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 orderings (Table 7). Here, GOLD refers to the gold-standard ordering for the 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 created the better result by raising the ordering of Newton. Secondly, BASE generated a bad ordering for '*baked goods (delicious)*' (KANTO (ALL)) but the resulting orderings of our method correlated with the gold-standard ordering by exploiting the expanded adjectives and general domain information. Thirdly, as for '*flesh (expensive)*' (KINKI (MALE)), BASE did not create a convincing ordering because the small amount of data was available 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 solve 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 in more specific domains where a small amount of text is available by exploiting: 1) the adjectives whose intensity correlates with that of target adjectives and 2) the global information of more general domains than a 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 orderings and the gold-standard orderings. Addressing the data sparseness problem, this paper opened a way to inferring values in more specific domains, or ultimately, individual values. We confirmed that we need more improvements through the evaluation by combining smoothing techniques.

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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*, Osaka, Japan, 2016.

[6] 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*, 2016.

[7] 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.

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

[9] Toshiharu Nishina, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 対義形容詞対との相互情報量を利用した概念語の順序付け. *IPSJ SIG NL*, 2013(8):1–7, Nov 2013 (in Japanese).

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

[11] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[12] 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*, pages 417–424, 2002.

[13] 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.

[14] 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.