Entity Set Expansion using Interactive Topic Information

Kugatsu Sadamitsu, Kuniko Saito, Kenji Imamura and **Yoshihiro Matsuo**
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
{sadamitsu.kugatsu, saito.kuniko, imamura.kenji, matsuo.yoshihiro}
@lab.ntt.co.jp

Abstract

We propose a new method for entity set expansion that achieves highly accurate extraction by suppressing the effect of semantic drift; it requires a small amount of interactive information. We supplement interactive information to re-train the topic models (based on interactive Unigram Mixtures) not only the contextual information. Although the topic information extracted from an unsupervised corpus is effective for reducing the effect of semantic drift, the topic models and target entities sometimes suffer grain mismatch. Interactive Unigram Mixtures can, with very few interactive words, ease the mismatch between topic and target entities. We incorporate the interactive topic information into a two-stage discriminative system for stable set expansion. Experiments confirm that the proposal raises the accuracy of the set expansion system from the baselines examined.

1 Introduction

The task of this paper is entity set expansion in which the lexicons are expanded from just a few seed entities (Pantel et al., 2009). For example, the user inputs the words "Apple", "Google" and "IBM", and the system outputs "Microsoft", "Facebook" and "Intel". Many set expansion and relation extraction algorithms are based on bootstrapping algorithms (Thelen and Riloff, 2002; Pantel and Pennacchiotti, 2006), which iteratively acquire new entities from corpora. These algorithms suffer from the general problem of "semantic drift". Semantic drift moves the extraction criteria away from the initial criteria demanded by the user and so reduces the accuracy of extraction.

Recently, topic information is being used to alleviate semantic drift. Topic information means the genre of each document as estimated by statistical topic models. Sadamitsu et al. (2011) proposed a bootstrapping method that uses unsupervised topic information estimated by Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to alleviate semantic drift. They use a discriminative method (Bellare et al., 2006) in order to incorporates topic information. They showed that the use of topic information improves the accuracy of the extracted entities.

Although unsupervised topic information has been confirmed to be effective, the topic models and target entity sometimes demonstrate grain mismatch. To avoid this mismatch, we refine the topic models to match the target entity grain. Deciding the entity grain from only positive seeds is difficult (Vyas et al., 2009). For example, the positive seed words are "Prius" and "Civic". In this situation, whether "Cadillac" is positive or negative depends on the user's definition. If the user thinks that "*Japanese car*" is positive grain, "Cadillac" should be placed into the negative class but if "*car*" is the positive grain it should be placed into the positive class. Note that we use the term "class" to refer to a set of entities denoted as C_P.

We control the topic models using not only positive seed entities but also a very small number of negative entities as distinguished from the output of the preliminary set expansion system. To implement this approach, we need topic models that offer con-

26th Pacific Asia Conference on Language,Information and Computation pages 108–116

trollability through the addition of negative words and high response speed for re-training. We utilize a variation of interactive topic models: interactive Unigram Mixtures (Sadamitsu et al., 2012). In a later section, we show that proposed method improves the accuracy of a set expansion system.

2 Set expansion using Topic information

2.1 Basic bootstrapping methods with discriminative models

In this section, we describe the basic method adopted from Bellare et al. (2006) since it offers easy handling of arbitrary features including topic information. At first, N_{ent} positive seed entities and N_{attr} seed attributes are given. The set of positive entity-attribute tuple, T_P, is obtained by taking the cross product of seed entity lists and attribute lists. Tuples T_P are used as queries for retrieving some documents, those that include a tuple present in T_P. Document set $D_{ent,attr}$ that includes the tuple $\{ent, attr\}$ is merged as an example to alleviate the sparseness of features.

Candidate entities are restricted to just the named entities that lie in close proximity to the seed attributes. Discriminative models are used to calculate the discriminative positive score, $s(ent, attr)$, of each candidate tuple, $\{ent, attr\}$. Their system extracts N_{new} new entities with high scores at each iteration as defined by the summation of $s(ent, attr)$ for all seed attributes (A_P); the condition is

$$\sum_{attr \in A_P} s(ent, attr) > 0. \quad (1)$$

Note that we do not iteratively extract new attributes because our purpose is entity set expansion.

2.2 Bootstrapping with Topic information

The discriminative approach is useful for handling arbitrary features. Although the context features and attributes partly reduce entity word sense ambiguity, some ambiguous entities remain. For example, consider the class "*car*" with the attribute of "*new model*". A false example is shown here: "A *new model* of *Android* will be released soon. The attractive smartphone begins to target new users who are ordinary people." The entity "*Android*" belongs to the "*cell-phone*" class, not "*car*", but appears with seed attributes or contexts because many "*cell-phones*" are introduced in "*new model*" as occurs with "*car*". By using topic, i.e. the genre of the document, we can distinguish "*Android*" from "*car*" and remove such false examples even if the false entity appeared with positive context strings or attributes.

Sadamitsu et al. (2011), the most relevant work to our current study, can disambiguate entity word senses and alleviate semantic drift by extracting topic information from LDA and adding it as discriminative features. The topic models can calculate the posterior probability $p(z|d)$ of topic z in document d. For example, the topic models give high probability to topic z ="*cell-phone*" in the above example sentences[1]. This posterior probability is effective for discrimination and is easily treated as a global feature of discriminative models. The topic feature value $\phi_t(z, ent, attr)$ is calculated as follows,

$$\phi_t(z, ent, attr) \propto \sum_{d \in D_{ent,attr}} p(z|d). \quad (2)$$

They also use topic information for selecting negative examples which are chosen far from the positive examples according to the measure of topic similarity.

There are other similar works. Paşca and Durme (2008) proposed clustering methods that are effective in terms of extraction, even though their clustering target is only the surrounding context. Ritter and Etzioni (2010) proposed a generative approach to allow extended LDA to model selection preferences. Although their approach is effective, we adopt the discriminative approach and so can treat arbitrary features including interactive information; moreover, it is applicable to bootstrapping methods.

3 Set expansion using Interactive Topic Information

3.1 Interactive Topic Information

Although topic information is effective for alleviating semantic drift, unsupervised topic information raises several problems. For example, Sadamitsu et al. (2011) reported that their set expansion system reached only 50% in the fine grained class "*car*";

[1] z is a random variable whose sample space is represented as a discrete variable, not explicit words.

an analysis showed that the nearest topic was mixed with "*motorcycle*". These classes are hard to distinguish even when both context and topic information are used simultaneously because they have similar context and topic information. One reason for the ineffectiveness of topic information is that the topics in topic models have grain sizes that are inappropriate for the target class in set expansion. Even when we use seed entities for modeling the semi-supervised topic models, as in (Andrzejewski et al., 2009), estimating the appropriate grain size is difficult because of a lack of information about other topics and contra-examples.

In order to control grain size in topic models, this section introduces interactive topic models that permit free control via human interaction. This interaction also includes some negative examples which are very effective in modifying the topic models. Topic model modification is now possible with the recent proposal of the Interactive Topic model (ITM) (Hu and Boyd-graber, 2011), which is based on LDA with the Dirichlet Forest prior (Andrzejewski et al., 2009). ITM makes it possible to accept the alterations input by users and to revise the topic model accordingly. Although ITM can modify a topic model, the calculation cost is high because it uses Gibbs sampling. The factor of processing overhead is very important because the user must wait for system feedback before interaction is possible. If user-interactivity is to be well accepted, we need to raise the response speed.

3.2 Interactive Unigram Mixtures

To obtain faster response, we utilize interactive Unigram Mixtures (IUMs) (Sadamitsu et al., 2012). This section details IUMs. IUMs are based on the simplest topic model, Unigram Mixtures (UMs) (Nigam et al., 2000) which are defined as

$$p(D) = \prod_{d=1}^{D} \sum_{z} p(z) \prod_{v} p(v|z)^{n(v,d)}, \qquad (3)$$

where D is a set of documents, d a document, z a hidden topic of a document, v is word type, $n(v,d)$ is the word count of v in document d. $p(z)$ and $p(v|z)$ are the model parameters of UMs. Their approach is to use the standard EM algorithm to estimate UMs. The estimation is achieved by computing the following formulae,

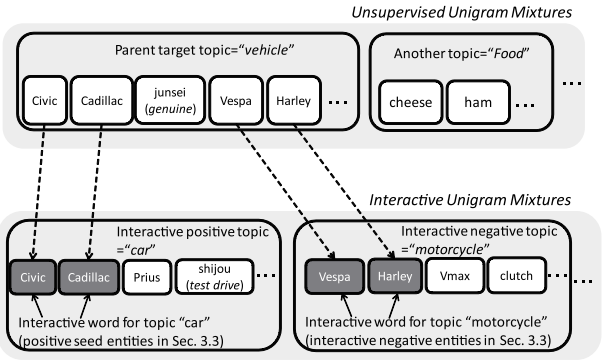

Figure 1: The abstract of interactive Unigram Mixtures with their characteristic topic words. The words in colored boxes are supervised words and the words in white boxes are the characteristic words extracted automatically. Note that, some characteristic topic words are not entity words.

$$p(v|z) = \frac{\sum_d n(v,d) p(z|d)}{\sum_v \sum_d n(v,d) p(z|d)} \qquad (4)$$

$$p(z) = \frac{\sum_d p(z|d)}{|D|}, \qquad (5)$$

where $p(z|d)$ is called the posterior probability of topic z for document d. For UMs, posterior probability $p(z|d)$ is calculated in E-step by the following formula,

$$p(z|d) = \frac{p(z) \prod_v p(v|z)^{n(v,d)}}{\sum_z p(z) \prod_v p(v|z)^{n(v,d)}}. \qquad (6)$$

UMs are not only faster than Gibbs sampling because only the standard EM algorithm is used, but they also make it easy to employ parallel processing (e.g. Map-Reduce).

IUMs are extended UMs and control each topic by utilizing a small set of interactive supervised words. Interactive updating involves using the interactive supervised words to re-model target topics as the set of child topics; for example, the interactive supervised words {Harley, Vespa} and {Civic, Cadillac} are used in order to re-model the target parent topic "*vehicle*" and construct the child topics "*motorcycle*" and "*car*", respectively, as shown in Figure 1. Note that, the words in white boxes in Figure 1 are example of characteristic topic words extracted by a score function such as $p(v|t)/p_{uni}(v)$, where

$p_{uni}(v)$ is a unigram model parameter for all documents. Note that, some characteristic topic words are not entity words because topic models describe all of words not only entity words (e.g. *"clutch"* in the "*motorcycle*" class).

In IUMs, we can focus on just a single parent topic which includes a subset of all documents e.g. *vehicle*. After creating unsupervised UMs, each document is clustered in topic z if its posterior probability satisfies $p(z|d) \geq 0.5$. Most documents meet this condition because UMs are uni-topic models unlike LDA, which offers multi-topic models. IUMs can be updated faster by this hard constraint because they process only the subset of documents.

In order to construct controlled topic models using very few supervised words, IUMs use supervised posterior probability $p_s(z|d_s)$. $p_s(z|d_s)$ is the probability of topic z according to document d_s that includes supervised words and is calculated as

$$p_s(z|d_s) = \frac{n_{d_s}(z)}{N_{d_s}}, \quad (7)$$

where $n_{d_s}(z)$ is the number of supervised words in document d_s that belong to topic z. N_{d_s} is the number of supervised words that belong to any topic, $N_{d_s} = \sum_z n_{d_s}(z)$. $p_s(z|d_s)$ is used instead of the E-step in estimating UMs (Eq. 6). For example, we consider two documents, $\{Civic, Cadillac\} \in d_{s1}$ and $\{Civic, Vespa\} \in d_{s2}$. The supervised posterior probability of d_{s1} and d_{s2} is calculated as $p_s(z = \text{"}\textit{Car}''|d_{s1}) = 1$ and $p_s(z = \text{"}\textit{Car}''|d_{s2}) = 0.5$, $p_s(z = \text{"}\textit{Motorcycle}''|d_{s2}) = 0.5$, respectively. These hypotheses can expand the supervised information from the word level to the document level.

The supervised posterior probability, $p_s(z|d_s)$, is too radical to be believed completely, so it is interpolated from the calculated posterior probabilities by the standard E-step in later iterations in the EM algorithms. The interpolated posterior probability $p_i(z|d_s)$ is calculated as

$$p_i(z|d_s) = w \cdot p_s(z|d_s) + (1 - w) \cdot p_c(z|d_s). \quad (8)$$

In the initial EM iteration, the interpolation weight w is set to 1, which means that only the supervised posterior probability is used. Interpolation weight w is decreased with each iteration. In early iterations, w takes a high value to permit model learning to closely approach the supervised structure. In later iterations, w is given a low value to adjust the total balance of model parameters from the perspective of probabilistic adequacy.

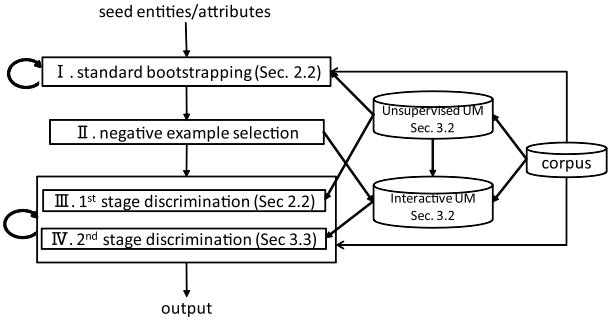

Figure 2: The structure of our system.

We note that the initial parameters are very important for modeling interactive topics appropriately. If the initial parameters are given at random, the model might converge on an inadequate local minima. To avoid this, the initial parameters are set to the parent topic model parameters.

3.3 Applying interactive Unigram Mixtures to set expansion

In this section we describe how to apply IUMs to set expansion in agreement with user's intuition. Our system's diagram is shown in Figure 2.

After the preliminary standard set expansion ("Ⅰ" in Figure 2) outputs some entities, we can choose interactive negative entities "E_{IN}" (e.g. "Harley, Vespa" in previous sections) found by either automatic methods (McIntosh and Curran, 2009) or manual selection ("Ⅱ" in Figure 2). Because this paper focuses on interactive control, it is out of scope as to which approach, automatic method or manual selection, should be used. In this paper, we choose few negative entities manually (in our experiments, we select two entities for each negative class). We choose not only E_{IN} but also their class names "C_{IN}" (e.g. "*motorcycle*" in previous sections) and treat them as negative "attributes" in the same way as seed attributes. IUMs are modeled using very little interactive information (E_{IN}, C_{IN}) as well as initial positive seed entities and attributes (E_P, A_P) as the supervised words for each child topic of target parent topic z_p. The target parent topic z_p is the one that

Table 1: Seed entities and seed attributes. The words surrounded by bracket are translation English. The words without brancket are appeared in Katakana or English itself.

class	seed entities	seed attributes
Car	Civic, Swift, Vitz, Corolla, Fit, Lexus, That's, Wagon R, Passo, Demio	kuruma (*car*), CM, shashu (*car line*),shinsha (*new car*), nosha (*delivering a car*), shingata (*new model*), engine, sedan, bumper, shaken (*automobile inspection*)
Dorama	Kita no Kuni kara, Tokugawa Yoshinobu, Mito Koumon, Nodame Cantabile, Dragon Sakura, Hana yori Dango, Furuhata Ninzaburo, ROOKIES, Aibou, Asunaro hakusho	dorama, meisaku (*master piece*), sakuhin (*product*), zokuhen (*sequel*), kantoku (*director*), shuen (*leader actor*), shutsuen (*appearance*), getsu-9 (*dorama started by Monday 9PM*), shichouritsu (*audience rate*), rendora (*miniseries*)
Soccer	Urawa Red Diamonds, Verdi, Avispa Fukuoka, Yokohama F Marinos, Barcelona, Real Madrid, Intel, Rome, Liverpool	soccer, J-League (*soccer league in Japan*), 1-bu (*Division 1*),goal

gives the highest $score(z)$,

$$z_p = \arg\max_z score(z), \quad (9)$$

$$score(z) = \sum_{v \in E_P} \frac{p(z)p(v|z)}{\sum_{z'} p(z')p(v|z')}, \quad (10)$$

where $p(z), p(v|z)$ are unsupervised UMs model parameters. Finally, the posterior probability calculated by IUMs is used as topic features as per the description in Sec 2.2.

Also we utilizes interactive negative entities not only for re-estimating the topic model but also for training the discriminative models as negative examples. Since there are only few interactive negative entities, we expand them by assuming that an entity co-occurring with an interactive negative class (C_{IN}) can be taken as negative entity "$E_{IN'}$". To summarize, interactive negative entity-attribute tuples "T_{IN}" are defined as in

$$T_{IN} = E_{IN} \times (C_{IN} + A_P) + E_{IN'} \times C_{IN},$$

where $\times$ indicates cross product. T_{IN} and T_P (described in Sec.2.1) are used as training data for discriminative models, negative and positive examples, respectively.

For using interactive information effectively, we adapt two stage discrimination. The first stage is the same as the original set expansion system with unsupervised topic model (described in Sec. 2.2); it achieves coarse grain general selection ("Ⅲ " in Figure 2). In the second stage, the system trains a discriminative model using the same number of positive and negative tuples selected from T_P and T_{IN} respectively with interactive topic information calculated by IUMs ("Ⅳ" in Figure 2). The system uses the trained discriminative model in the second stage to re-score the selected candidates from the first stage.

Although the single step discriminative approach can be utilized by using T_{IN} in the first stage as the supervised data, this would degrade discrimination performance. The discriminative models could not train fine and coarse grain simultaneously as same as UMs. In preliminary experiments on the one stage method, we confirmed that the system outputs many inadequate entities belonging to wrong topics in the sense of coarse grain.

McIntosh (2010) proposed the method most similar to ours. In McIntosh (2010), only negative entities are clustered based on distributional similarity. We cluster not only the entities themselves but also their topic information.

Vyas and Pantel proposed an interactive method for entities refinement and improved accuracy of set expansion (Vyas and Pantel, 2009). They utilized the similarity method (SIM) and feature modification method (FMM) for refinement of entities and their local context features.

As far as we know, our proposal represents the first interactive method designed for the set expansion task with topic information. By incorporating interactive topic information, we can expect that the accuracy is improved since an improvement is achieved with unsupervised topic information.

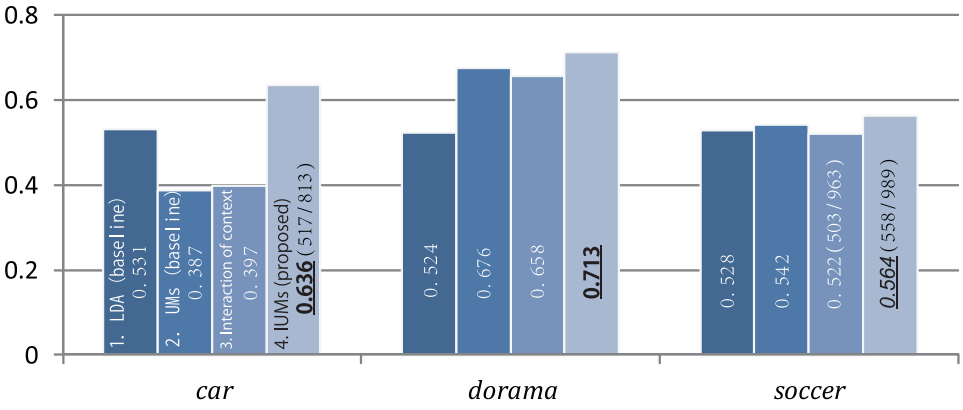

Figure 3: Results for the three classes "*car*", "*dorama*" and "*soccer team*". Bold font indicates that the difference in accuracy between proposal and best of baseline is significant by binomial test with $P < 0.05$ and italic font indicates $P < 0.1$.

4 Experiments

4.1 Experimental Settings

The experimental parameters follow those of the experiments in Sadamitsu et al. (2011). We used 30M Japanese blog articles crawled in May 2008. The documents were tokenized, chunked, and labeled with IREX 8 named entity types (Fuchi and Takagi, 1998; Suzuki et al., 2006), and transformed into context features. The context features were defined using the template "(head) *ent.* (mid.) *attr.* (tail)". The words included in each part were used as surface, part-of-speech, and named entity label features with added position information. Maximum word number of each part was set at 2. The features have to appear in both the positive and negative training data at least 5 times.

In the experiments, we used three classes, "*car*", "*dorama*" and "*soccer team*" since they often suffer semantic drift. The adjustment numbers for the basic setting are $N_{ent} = 10, N_{attr} = 10, N_{new} = 100, |C_{IN}| = 2$. Note that, for confirmation in a more severe situation, we set $N_{attr} = 4, |C_{IN}| = 1$ in "*soccer*" class. After running 10 Bootstrapping iterations, we obtained 1000 entities in total. The seed entities and attributes for each class are shown in Table 1

SVM^{light} (Joachims, 1999) with a second order polynomial kernel was used as the discriminative model. Unsupervised UMs and unsupervised LDA were used for training 100 mixture topic models. Parallel LDA, which is LDA with MPI (Liu et al., 2011), was used for training and inference for LDA. For training IUMs, we set the mixture number of child topics to 5, that covers both interactive and other unsupervised topics about each class. The other unsupervised topics, $(5 - (|C_{IN}| + |C_P|))$, catch the other structure in the parent topic z_p, where $|C_P|$ always equal to 1.

Four settings were examined.

- First is a baseline method using unsupervised topic information with LDA (without interaction); it is described in Sec. 2.2.
- Second is similar to first but the topic models, LDA, are replaced by unsupervised UMs.
- Third is the second setting with the addition of the set of interactive tuples, T_{IN}, for re-training discriminative models using only context information. This setting allows confirmation of just the IUMs effect by comparison to fourth setting which also models interactive topic information.
- Fourth, proposed, is the third setting with the addition of the IUMs learned from the set of interactive tuples, T_{IN}.

Each extracted entity is labeled with *correct* or *incorrect* by two evaluators based on the results of a commercial search engine. Some of the results

Table 2: Examples of extracted entities (first column) and characteristic topic words extracted from UMs and IUMs (fourth column). This table also shows interactive supervised positive and negative classes (second column) and their supervised entities (third column). The words with underline are incorrect extracted entity in the first column and incorrect characteristic topic words in the fourth column.

Extracted entities baseline(UM)&proposed(IUM)	Interactive classes (C_P&C_{IN})	Interactive entities (E_P&E_{IN})	Extracted topic words from each topic
Class = "car"	parent posi. (z_p)	*(=seed entities)*	tosou (*paint*), secchaku (*bond*), plug, junsei (*genuine*)
baseline: Sylvia, Harley, E700	interactive posi. C_P=*car*	*(=seed entities)*	turbo, kuruma (*car*), wheel, shijou (*test drive*)
proposed: Sylvia, 117 coupe, nubi250 (car navigation system)	interactive nega.1 C_{IN_1}=*motorcycle*	Harley, CB400	baiku (*motorcycle*), plug, bolt, clutch
	interactive nega.2 C_{IN_2}=*train*	E700, E531 (*train names*)	kado (*movable*), ganpura (*plamodel of robot*), puramo (*plamodel*), Bandai (*plamodel company*)
Class = "dorama"	parent posi. (z_p)	*(=seed entities)*	Juri Ueno, Masami Nagasawa, (actoresses), Cannes, Hachiwan Diver (*anime title*)
baseline: Prison Break, Iron Man, Konan	interactive posi. C_P="*dorama*"	*(=seed entities)*	Juri Ueno, Masami Nagasawa, Last Friends (*dorama title*), shichouritsu (*viewer rate*)
proposed: Prison Break, Shinsengumi!, Tokudane! (news program)	interactive nega.1 C_{IN_1}= *movie*	Kung Fu Panda Iron Man	Cannes, Masami Nagasawa, Akunin (*movie title*), shishakai (*preview*)
	interactive nega.2 C_{IN_2}="*anime*"	Konan, Negima (*anime titles*)	TV Tokyo (*broadcasting many animes*), OVA (*original video anime*), Oricon, Yatta-man (*anime title*)
Class = "soccer"	parent posi. (z_p)	(=*seed entities*)	Chelsea, toushu (*pitcher*), anda (*hit*), shitten (*loss a point*)
baseline: A Madrid, Giants	interactive posi. C_P="*soccer*"	(=*seed entities*)	Manchester United, DF, FW, FC Tokyo (soccer team name)
proposed: A. Madrid, Manchester C, Football Association (not team)	interactive nega. C_{IN}="*baseball*"	Giants, Tigers (*baseball teams*)	anda (*hit*), toushu (*pitcher*), kai omote (*top of*), shikyuu (*ball four*)

(1231 entities) were double checked and the κ score for agreement between evaluators was 0.843.

4.2 Results

Figure 3 compares the accuracy of the four methods. If the number of extracted examples is lower than 1000, i.e. Eq. 1 was unsatisfied, the figure shows the number of extracted examples and the correct number in brackets. At first, we compare two baseline methods, first and second bar, that use different unsupervised topic models. The result is that "UMs" are superior to "LDA" in "*dorama*" but inferior in "*car*". They yield more variability than "LDA". One reason for this is that UMs are uni-topic models which leads to over-fitting. Uni-topic models describe most documents by one topic. For uni-topic models, setting a small number of topics (topic grain size is large) suits large topics rather than than small topics because the latter would have to be merged to match the grain size. Conversely, setting a large number of topics suits small topics rather than large topics because the latter would have to split. This restriction can degrade accuracy significantly. LDA smoothes the topics due to its multi-topic modeling. The third setting shows that the interactive tuples T_{IN} used for re-modeling with only context information is not effective. We consider this result indi-

cates that context is not effective in terms of discrimination with fine grain, because at this grain positive context is similar to negative context. Proposed, on the other hand, offers improved accuracy in all classes significantly. These results show the effectiveness of the interactive method that uses topic information. The interactive methods are more effective than the selection of topic model type.

To confirm whether our proposal works properly, we show characteristic topic words extracted from IUMs with interactive classes (C_P, C_{IN}) and entities (E_P, E_{IN}) in Table 2. Because each topic z is not explicitly understandable, we use the characteristic topic words which are representative words for each topic z. The characteristic topic words are ranked by a score function $p(v|t)/p_{uni}(v)$.

- The first column shows target classes and the resulting entities yielded by using set expansion of baseline with UM and proposed method with IUM.
- The second column shows the parent positive topic (z_p) selected by Eq.(9), seed class (C_P) and the interactive supervised classes (C_{IN}) as interactive topic information.
- The third column shows the seed entities (E_P) and the interactive supervised negative entities (E_{IN}).
- The fourth column shows the characteristic topic words of each topic. In this experiment, we extracted 4 topic words from the words listed in top 10.

Table 2 shows that the characteristic topic words are strongly related to the interactive positive (negative) classes and their entities. For example, in the parent positive topic of "*dorama*" class in Figure 2, there are some characteristic topic words, "Juri Ueno", "Masami Nagasawa" (*actresses*), "Cannes" and "Hachiwan Diver (*anime title*)". The words with underline are inadequate topic words for "*dorama*" class. After applying IUM, in the interactive positive topic, the topic words are refined as adequate words, "shichouritsu (*viewer rate*)" and a dorama title. IUMs also model appropriately for the interactive negative topic "*movie*" whose extracted topic words are "Cannes" and "shishakai (*preview*)".

On the other hand, in the "*motorcycle*" class which is the first interactive negative class for "*car*" class, topic words include "plug", "bolt" and "clutch". Although these words are not uniquely "*motorcycle*" words, they tend to appear with "*motorcycle*" class in the corpus used. There are many inadequate characteristic topic words extracted for the "*train*" class, which is the second negative class of the "*car*" class. The characteristic topic words are placed into the "*plamodel*" (plastic model) topic. We consider that the "*train*" words were extracted by the "*plamodel*" topic via semantic drift. This situation is assumed as an example of human's mispredication for a negative topic. Even if IUMs model a class (*plamodel*) different from user prediction topic (*train*), interactive topic information is also effective for alleviating semantic drift. As a result, "*car*" class as the interactive positive topic, its topic words are more pure like "turbo" and "shijou (*test drive*)" than in the parent positive topic.

A similar observation is confirmed from the "*soccer*" class. Because the interactive negative information is smaller than other classes, the improvement of accuracy is smaller. We can expect that much more interactive information achieve further improvement for the accuracy.

5 Conclusion

We proposed an approach to set expansion that uses interactive information for refining the topic model and showed that it can improve expansion accuracy. In our set expansion system, 2 stage discriminations are applied to discriminate coarse from fine grain in each stage. Since we also applied interactive Unigram Mixtures for treating interactive information, our set expansion system makes interaction highly effective.

The remaining problem is how to automatically determine the most appropriate threshold in set expansion. Also, we intend to compare the effectiveness of using manually detected negative examples (which were used in this paper) and automatically detected interactive negative examples.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating Domain Knowledge into Topic

Modeling via Dirichlet Forest Priors. In *Proceedings of the International Conference on Machine Learning.*, volume 382, pages 25–32.

Kedar Bellare, Partha P. Talukdar, Giridhar Kumaran, Fernando Pereira, Mark Liberman, Andrew McCallum, and Mark Dredze. 2006. Lightly-supervised attribute extraction. In *Proceedings of the Advances in Neural Information Processing Systems Workshop on Machine Learning for Web Search*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Takeshi Fuchi and Shinichiro Takagi. 1998. Japanese Morphological Analyzer using Word Co-occurrence-JTAG. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 409–413.

Yuening Hu and Jordan Boyd-graber. 2011. Interactive Topic Modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 248–257.

Thorsten Joachims. 1999. *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*. Software available at `http://svmlight.joachims.org/`.

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. PLDA+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology, special issue on Large Scale Machine Learning*. Software available at `http://code.google.com/p/plda`.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 396–404.

Tara McIntosh. 2010. Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 356–365.

Kamal Nigam, Andrew K Mccallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134.

Marius Paşca and Benjamin Van Durme. 2008. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 19–27.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 938–947.

Alan Ritter and Oren Etzioni. 2010. A Latent Dirichlet Allocation method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Kugatsu Sadamitsu, Kuniko Saito, Kenji Imamura, and Genichiro Kikui. 2011. Entity Set Expansion using Topic information. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 726–731.

Kugatsu Sadamitsu, Kuniko Saito, Kenji Imamura, and Yoshihiro Matsuo. 2012. Constructing a Class-Based Lexical Dictionary using Interactive Topic Models. In *Proceedings of the 8th International Language Resources and Evaluation*, pages 2590–2595.

Jun Suzuki, Erik McDermott, and Hideki Isozaki. 2006. Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 217–224.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 conference on Empirical methods in natural language processing*, pages 214–221.

Vishnu Vyas and Patrick Pantel. 2009. Semi-automatic entity set refinement. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 290–298.

Vishnu Vyas, Patrick Pantel, and Eric Crestan. 2009. Helping editors choose better seed sets for entity set expansion. In *Proceeding of the 18th ACM conference on Information and Knowledge Management*, pages 225–234.