Entity Set Expansion using Topic information

Kugatsu Sadamitsu, Kuniko Saito, Kenji Imamura and **Genichiro Kikui**[*]
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
{sadamitsu.kugatsu, saito.kuniko, imamura.kenji}@lab.ntt.co.jp
kikui@cse.oka-pu.ac.jp

Abstract

This paper proposes three modules based on latent topics of documents for alleviating "semantic drift" in bootstrapping entity set expansion. These new modules are added to a discriminative bootstrapping algorithm to realize topic feature generation, negative example selection and entity candidate pruning. In this study, we model latent topics with LDA (Latent Dirichlet Allocation) in an unsupervised way. Experiments show that the accuracy of the extracted entities is improved by 6.7 to 28.2% depending on the domain.

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 a few words "Apple", "Google" and "IBM" , and the system outputs "Microsoft", "Facebook" and "Intel".

Many set expansion algorithms are based on bootstrapping algorithms, which iteratively acquire new entities. 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. Pantel and Pennacchiotti (2006) proposed Espresso, a relation extraction method based on the co-training bootstrapping algorithm with entities and attributes. Espresso alleviates semantic-drift by a sophisticated scoring system based on pointwise mutual information (PMI). Thelen and Riloff (2002), Ghahramani and Heller (2005) and Sarmento et al. (2007) also proposed original score functions with the goal of reducing semantic-drift.

Our purpose is also to reduce semantic drift. For achieving this goal, we use a discriminative method instead of a scoring function and incorporate topic information into it. Topic information means the genre of each document as estimated by statistical topic models. In this paper, we effectively utilize topic information in three modules: the first generates the features of the discriminative models; the second selects negative examples; the third prunes incorrect examples from candidate examples for new entities. Our experiments show that the proposal improves the accuracy of the extracted entities.

The remainder of this paper is organized as follows. In Section 2, we illustrate discriminative bootstrapping algorithms and describe their problems. Our proposal is described in Section 3 and experimental results are shown in Section 4. Related works are described in Section 5. Finally, Section 6 provides our conclusion and describes future works.

2 Problems of the previous Discriminative Bootstrapping method

Some previous works introduced discriminative methods based on the logistic sigmoid classifier, which can utilize arbitrary features for the relation extraction task instead of a scoring function such as Espresso (Bellare et al., 2006; Mintz et al., 2009). Bellare et al. reported that the discriminative approach achieves better accuracy than Espresso when the number of extracted pairs is increased because

[*] Presently with Okayama Prefectural University

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 726–731,
Portland, Oregon, June 19-24, 2011.

multiple features are used to support the evidence.

However, three problems exist in their methods. First, they use only local context features. The discriminative approach is useful for using arbitrary features, however, they did not identify which feature or features are effective for the methods. Although the context features and attributes partly reduce entity word sense ambiguity, some ambiguous entities remain. For example, consider the domain *broadcast program* (PRG) and assume that PRG's attribute is *advertisement*. A false example is shown here: "*Android* 's *advertisement* employs Japanese popular actors. The attractive smartphone begins to target new users who are ordinary people." The entity *Android* belongs to the *cell-phone* domain, not PRG, but appears with positive attributes or contexts because many *cell-phones* are introduced in *advertisements* as same as *broadcast program*. By using topic, i.e. the genre of the document, we can distinguish "Android" from PRG and remove such false examples even if the false entity appeared with positive context strings or attributes. Second, they did not solve the problem of negative example selection. Because negative examples are necessary for discriminative training, they used all remaining examples, other than positive examples, as negative examples. Although this is the simplest technique, it is impossible to use all of the examples provided by a large-scale corpus for discriminative training. Third, their methods discriminate all candidates for new entities. This principle increases the risk of generating many false-positive examples and is inefficient. We solve these three problems by using topic information.

3 Set expansion using Topic information

3.1 Basic bootstrapping methods

In this section, we describe the basic method adopted from Bellare (Bellare et al., 2006). Our system's configuration diagram is shown in Figure 1. In Figure 1, arrows with solid lines indicate the basic process described in this section. The other parts are described in the following sections. After N_s positive seed entities are manually given, every noun co-occurring with the seed entities is ranked by PMI scores and then selected manually as N_a positive attributes. N_s and N_a are predefined basic adjustment numbers. The entity-attribute pairs are obtained by taking the cross product of seed entity lists and attribute lists. The pairs are used as queries for retrieving the positive documents, which include positive pairs. The document set $D_{e,a}$ including same entity-attribute pair $\{e, a\}$ is regarded as one example $E_{e,a}$ to alleviate over-fitting for context features. These are called positive examples in Figure 1. Once positive examples are constructed, discriminative models can be trained by randomly selecting negative examples.

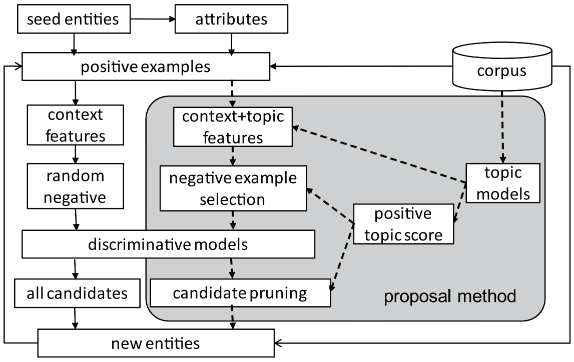

Figure 1: The structure of our system.

Candidate entities are restricted to only the Named Entities that lie in the close proximity to the positive attributes. These candidates of documents, including Named Entity and positive attribute pairs, are regarded as one example the same as the training data. The discriminative models are used to calculate the discriminative positive score, $s(e, a)$, of each candidate pair, $\{e, a\}$. Our system extracts N_n types of new entities with high scores at each iteration as defined by the summation of $s(e, a)$ of all positive attributes (A_P); $\sum_{a \in A_P} s(e, a)$. Note that we do not iteratively extract new attributes because our purpose is entity set expansion.

3.2 Topic features and Topic models

In previous studies, context information is only used as the features of discriminative models as we described in Section 2. Our method utilizes not only context features but also topic features. By utilizing topic information, our method can disambiguate the entity word sense and alleviate semantic drift. In order to derive the topic information, we utilize statistical topic models, which represent the relation between documents and words through hidden topics. 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 useful as a global feature for discrimination. The topic feature value $\phi_t(z, e, a)$ is calculated as follows.

$$\phi_t(z, e, a) = \frac{\sum_{d \in D_{e,a}} p(z|d)}{\sum_{z'} \sum_{d \in D_{e,a}} p(z'|d)}.$$

In this paper, we use Latent Dirichlet Allocation (LDA) as the topic models (Blei et al., 2003). LDA represents the latent topics of the documents and the co-occurrence between each topic.

In Figure 1, shaded part and the arrows with broken lines indicate our proposed method with its use of topic information including the following sections.

3.3 Negative example selection

If we choose negative examples randomly, such examples are harmful for discrimination because some examples include the same contexts or topics as the positive examples. By contrast, negative examples belonging to broad genres are needed to alleviate semantic drift. We use topic information to efficiently select such negative examples.

In our method, the negative examples are chosen far from the positive examples according to the measure of topic similarity. For calculating topic similarity, we use a ranking score called "positive topic score", $PT(z)$, defined as follows, $PT(z) = \sum_{d \in D_P} p(z|d)$, where D_P indicates the set of positive documents and $p(z|d)$ is topic posterior probability for a given positive document. The bottom 50% of the topics sorted in decreasing order of positive topic score are used as the negative topics. Our system picks up as many negative documents as there are positive documents with each selected negative topic being equally represented.

3.4 Candidate Pruning

Previous works discriminate all candidates for extracting new entities. Our basic system can constrain the candidate set by positive attributes, however, this is not enough as described in Section 2. Our candidate pruning module, described below, uses the measure of topic similarity to remove obviously incorrect documents.

This pruning module is similar to negative example selection described in the previous section. The positive topic score, PT, is used as a candidate constraint. Taking all positive examples, we select the positive topics, PZ, which including all topics z satisfying the condition $PT(z) > th$. At least one topic with the largest score is chosen as a positive topic when $PT(z) \leq th$ about all topics. After selecting this positive topic, the documents including entity candidates are removed if the posterior probability satisfy $p(z|d) \leq th$ for all topics z. In this paper, we set the threshold to $th = 0.2$. This constraint means that the topic of the document matches that of the positive entities and can be regarded as a hard constraint for topic features.

4 Experiments

4.1 Experimental Settings

We use 30M Japanese blog articles crawled in May 2008. The documents were tokenized by JTAG (Fuchi and Takagi, 1998), chunked, and labeled with IREX 8 Named Entity types by CRFs using Minimum Classification Error rate (Suzuki et al., 2006), and transformed into features. The context features were defined using the template "(head) *entity* (mid.) *attribute* (tail)". The words included in each part were used as surface, part-of-speech and Named Entity label features added position information. Maximum word number of each part was set at 2 words. The features have to appear in both the positive and negative training data at least 5 times.

In the experiments, we used three domains, car ("CAR"), broadcast program ("PRG") and sports organization ("SPT"). The adjustment numbers for basic settings are $N_s = 10, N_a = 10, N_n = 100$. After running 10 iterations, we obtained 1000 entities in total. SVM^{light} (Joachims, 1999) with second order polynomial kernel was used as the discriminative model. Parallel LDA, which is LDA with MPI (Liu et al., 2011), was used for training 100 mixture topic models and inference. Training corpus for topic models consisted of the content gathered from

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

	CAR	PRG	SPT
1. Baseline	0.249	0.717	0.781
2. Topic features + 1.	**0.483**	0.727	**0.844**
3. Negative selection + 2.	**0.509**	**0.762**	0.846
4. Candidate pruning + 3.	*0.531*	**0.824**	0.848

Table 1: The experimental results for the three domains. Bold font indicates that the difference between accuracy of the methods in the row and the previous row is significant ($P < 0.05$ by binomial test) and italic font indicates ($P < 0.1$).

14 days of blog articles. In the Markov-chain Monte Carlo (MCMC) method, sampling was iterated 200 times for training with a burn-in taking 50 iterations. These parameters were selected based on the results of a preliminary experiment.

Four experimental settings were examined. First is Baseline; it is described in Section 3.1. Second is the first method with the addition of topic features. Third is the second method with the addition of a negative example selection module. Fourth is the third method with the addition of a candidate pruning module (equals the entire shaded part in Figure 1). Each extracted entity is labeled with *correct* or *incorrect* by two evaluators based on the results of a commercial search engine. The κ score for agreement between evaluators was 0.895. Because the third evaluator checked the two evaluations and confirmed that the examples which were judged as correct by either one of the evaluators were correct, those examples were counted as correct.

4.2 Experimental Results

Table 1 shows the accuracy and significance for each domain. Using topic features significantly improves accuracy in the CAR and SPT domains. The negative example selection module improves accuracy in the CAR and PRG domains. This means the method could reduce the risk of selecting false-negative examples. Also, the candidate pruning method is effective for the CAR and PRG domains. The CAR domain has lower accuracy than the others. This is because similar entities such as motorcycles are extracted; they have not only the same context but also the same topic as the CAR domain. In the SPT domain, the method with topic features offer significant improvements in accuracy and no further improvement was achieved by the other two modules.

To confirm whether our modules work properly, we show some characteristic words belonging to each topic that is similar and not similar to target domain in Table 2. Table 2 shows characteristic words for one positive topic z_h and two negative topics z_l and z_e, defined as follow.

- z_h (the second row) is the topic that maximizes $PT(z)$, which is used as a positive topic.
- z_l (the fourth row) is the topic that minimizes $PT(z)$, which is used as a negative topic.
- z_e (the fifth row) is a topic that, we consider, effectively eliminates "drifted entities" extracted by the baseline method. z_e is eventually included in the lower half of topic list sorted by $PT(z)$.

For a given topic, z, we chose topmost three words in terms of topic-word score. The topic-word score of a word, v, is defined as $p(v|z)/p(v)$, where $p(v)$ is the unigram probability of v, which was estimated by maximum likelihood estimation. For utilizing candidate pruning, near topics including z_h must be similar to the domain. By contrast, for utilizing negative example selection, the lower half of topics, z_l, z_e and other negative topics, must be far from the domain. Our system succeeded in achieving this. As shown in "CAR" in Table 2, the nearest topic includes "shaken" (*automobile inspection*) and the farthest topic includes "naika" (*internal medicine*) which satisfies our expectation. Furthermore, the effective negative topic is similar to the topic of drifted entity sets (*digital device*). This indicates that our method successfully eliminated drifted entities. We can confirm that the other domains trend in the same direction as "CAR" domain.

5 Related Works

Some prior studies use every word in a document/sentence as the features, such as the distributional approaches (Pantel et al., 2009). These methods are regarded as using global information, however, the space of word features are sparse, even if the amount of data available is large. Our approach can avoid this problem by using topic models which

domain	CAR	PRG	SPT
words of the nearest topic z_h (highest PT score)	shaken (*automobile inspection*), nosha (*delivering a car*), daisha (*loaner car*)	Mari YAMADA, Tohru KUSANO, Reiko TOKITA (*Japanese stars*)	toshu (*pitcher*), senpatsu (*starting member*), shiai (*game*)
drifted entities (using baseline)	iPod, mac (*digital device*)	PS2, XBOX360 (*video game*)	B'z, CHAGE&ASKA (*music*)
words of effective negative topic z_e (Lower half of PT score)	gaso (*pixel*), kido (*brightness*), mazabodo (*mother board*)	Lv. (*level*), kariba (*hunting area*), girumen (*guild member*)	sinpu (*new release*), X JAPAN , Kazuyoshi Saito (*Japanese musicians*)
words of the farthest topic z_l (Lowest PT score)	naika (*internal medicine*), hairan (*ovulation*), shujii (*attending doctor*)	tsure (*hook a fish*), choka (*result of hooking*), choko (*diary of hooking*)	toritomento (*treatment*), keana (*pore*), hoshitsu (*moisture retention*)

Table 2: The characteristic words belonging to three topics, z_h, z_l and z_e. z_h is the nearest topic and z_l is the farthest topic for positive entity-attribute seed pairs. z_e is an effective negative topic for eliminating "drifted entities" extracted by the baseline system.

are clustering methods based on probabilistic measures. By contrast, 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 use extended LDA to model selectional preferences. Although their approach is similar to ours, our approach is discriminative and so can treat arbitrary features; it is applicable to bootstrapping methods.

The accurate selection of negative examples is a major problem for positive and unlabeled learning methods or general bootstrapping methods and some previous works have attempted to reach a solution (Liu et al., 2002; Li et al., 2010). However, their methods are hard to apply to the Bootstrapping algorithms because the positive seed set is too small to accurately select negative examples. Our method uses topic information to efficiently solve both the problem of extracting global information and the problem of selecting negative examples.

6 Conclusion

We proposed an approach to set expansion that uses topic information in three modules and showed that it can improve expansion accuracy. The remaining problem is that the grain size of topic models is not always the same as the target domain. To resolve this problem, we will incorporate the active learning or the distributional approaches. Also, comparisons with the previous works are remaining work. From another perspective, we are considering the use of graph-based approaches (Komachi et al., 2008) incorporated with the topic information using PHITS (Cohn and Chang, 2000), to further enhance entity extraction accuracy.

References

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.

David Cohn and Huau Chang. 2000. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*, pages 167–174.

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.

Zoubin Ghahramani and Katherine A. Heller. 2005. Bayesian sets. In *Proceedings of the Advances in Neural Information Processing Systems*.

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

Mamoru Komachi, Taku Kudo, Masashi Shimbo, and Yuji Matsumoto. 2008. Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1011–1020.

Xiao-Li Li, Bing Liu, and See-Kiong Ng. 2010. Negative Training Data can be Harmful to Text Classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 218–228.

Bing Liu, Wee S. Lee, Philip S. Yu, and Xiaoli Li. 2002. Partially supervised classification of text documents. In *Proceedings of the 19th International Conference on Machine Learning*, pages 387–394.

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`.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

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 ACL Conference*, pages 424–434.

Luis Sarmento, Valentin Jijkuon, Maarten de Rijke, and Eugenio Oliveira. 2007. More like these: growing entity classes from seeds. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 959–962.

Jun Suzuki, Erik McDermott, and Hideki Isozaki. 2006. Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the 21st COLING and 44th ACL Conference*, 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.