Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 22 (2013) 1311 – 1318

Procedia Computer Science

17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013

The Prediction of Ellipses Using Topic Model for Japanese Colloquial Inquiry Text

Tomohiko Harada[a,*], Yoshikatsu Fujita[b], Kazuhiko Tsuda[c]

[a]*Graduate School of Systems and Information Engineering, University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan*
[b]*Department of Sociology, Teikyo University, 359 Otsuka, Hachioji, Tokyo 192-0395, Japan*
[c]*Graduate School of Business Sciences University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan*

Abstract

Generally inquiries through Web forms and e-mails are increasing. These inquiry texts usually include many informal expressions use of the colloquial style and many omitted words. An omitted word causes the meaning of a sentence to become ambiguous and makes the reader misread and misunderstand a context. In this paper we propose a method to predict omitted words from context and knowledge using topic information. From the results of evaluation experiment, we have confirmed that some of our methods can predict omitted words at the accuracy rate more than 40% for the expression that we used in the experiment.

Keywords: Colloquial expressions; Ellipsis; Statistical topic models; Gibbs sampling; LDA

1. Introduction

Recently, opportunities to ask companies and local governments using web and e-mail about the contents of their e-mails and use of their service and products have increased. For them it is an opportunity to gain the trust of residents and consumers by using e-mail and web to carry out their service and answer queries promptly and appropriately. But failure to do so could cause the opposite and result in loss of the consumers' trust. For example, because of an inappropriate correspondence, a rumor on the internet may spread out which could cause negativity on their reputation. In this case they need to take damage control.

Usually, inquiry through web and email are in a plain text format, and there is no decoration on it. Recently, because of cellular phones and smart phones' keypad and small screen, people tend to omit longer sentences. In addition, it is not the literary style that is usually used in a letter or on a report. The informal and colloquial style expressions are used more and more because there are many users with a level of high anonymity and relatively young age. In particular, ellipses in an inquiries' text format can cause the oversight of an important question and misread or misunderstand. If speaking on the telephone, one can compensate for missing information through interaction, confirm the intent of the questioner, or revise the way of answers. But it is beyond the point of no return if you reply by e-mail. Therefore, understanding the query statement correctly is important.

*Corresponding author. Tel.: +81-3-3942-6869 ; fax: +81-3-3942-6829.
E-mail address: s1230165@u.tsukuba.ac.jp.

doi:10.1016/j.procs.2013.09.219

2. Characteristic of the Inquiry Text

Figure 1 shows an example of the inquiry text that is the subject of this study[1][2]. Figure 2 illustrates the inferences made inside the ellipsis. It is found in Figure 2 that many of ellipses shown in round parentheses in comparison with Figure1.

パソコンに DVD を入れてもディスクを読み取ってもらえなく、ドライブを見ても常に中身は空です。
しかし CD を再生することは可能です。
こういう場合は修理に出したほうがいいのでしょうか?
すみませんが、ご回答よろしくお願いします。

Fig. 1. Example of query text

(私 が)(私の) パソコン (のドライブ) に DVD(のディスク) を入れても、(このドライブ は) ディスク (のデータ) を読み取らず、(私 が)(エクスプローラで) ドライブ (の中) を見ても常に (ドライブの) 中身は空です。
しかし、(このパソコンの)(ドライブで)CD(のディスク) を再生することは可能です。
こういう場合は (私 は)(このパソコン を) 修理に出したほうがいいのでしょうか?
すみませんが、ご回答 (を) よろしくお願いします。

Fig. 2. Examples of some ellipsis complement

For example, it is found in Figure 2 that some of surface case[1] characterizing the meaning of the sentence, such as underlined ハ case (the subjective), ガ case (the nominative), ニ case (indirect object) and ヲ case (direct object) are omitted. In order to put this trend into perspective, we made a preliminary investigation that, using two kinds of data, counted the number of the appearances of four kinds of surface case per one sentence in 16,598 question texts[3]. One is question text data of "Yahoo! Chiebukuro data" in which a lot of colloquial style expressions are included in. The other is text data of the Japanese edition of "Wikipedia"[4] as the non-colloquial style text. The results are shown in Table 1. From the Table 1 we see that for "Yahoo! Chiebukuro data", the number of the appearances of the surface case remains approximately 50% in comparison with "Wikipedia". This shows the numerousness of the ellipsis in the colloquial inquiry text.

Table 1. Number of the appearances of the surface case per one sentence

Num. of cases	ハ (*ha*)	ガ (*ga*)	ニ (*ni*)	ヲ (*wo*)
Yahoo!Chiebukuro (colloquial)	0.274	0.347	0.263	0.385
Wikipdia (non-colloquial)	0.496	0.604	0.745	0.720

In this paper, we introduce some related research on an ellipsis resolution in Chapter 3. Then we report some methods to predict an omitted word, and an evaluation experiment and the result in Chapter 4.

3. Ellipses Resolution

In Japanese, an element of a sentence that can be predicted by context is often omitted. Furthermore, the element that has became the topic of the sentence is often omitted because a listener can still understand. In addition, many other kinds of elements are omitted[2].

[1] This is one of the questions that ware posted to the "Yahoo! Chiebukuro" which is a QA site typical in Japan.
[2] The data includes 16,257,413 questions and 50,053,894 answers recorded between 4/1/2004～4/7/2009 and offered to the researcher.
[3] We used 16,598 data in the subcategory of "PC" of "Yahoo! Chiebukuro data" of Mar 2009.
[4] We used 16,598 paragraphs of Japanese "Wikipedia" data, dated Aug 2012.

There is a study of anaphora resolution in research to supply the omitted word. The anaphora is that a phrase (antecedent) and a phrase (anaphora) point the same content in the same context. The ellipsis means anaphora by the zero pronouns. The anaphora analysis in particular is an important task of natural language processing because Japanese has many ellipsis.

3.1. Zero Anaphora Resolution

The study on anaphora resolution carried out flourishingly but much of it intended to be solved by surface cues that appear in a context. For example, by the centering theory[3], a topic remains in the "center" of the context, and the "center" supposes that it is easy to become an antecedent of anaphora and ellipsis. There is a study by Walker et al.[4] on research dealing with anaphora resolution in Japanese based on the centering theory. Walker et al. displayed the candidates of the antecedent from a manifestation-related height in order of "sub jective > nominative > indirect ob ject > direct object > and others" and proposed a method to decide the antecedent of zero pronouns from the ranking. However, the centering theory has the problem that it cannot resolve appropriately when there are some anaphora in a sentence and the above sentence does not have an antecedent.

In a recent study by Hayashibe et al.[5] on anaphora resolution, Hayashibe et al. focused on it cannot be determined that the ガ case of "自首した" is X with only the sentence "X を逮捕した" by the method of Walker et al. and the centering theory. Hayashibe et al. defined "the case particle + verb" as the case structure and suggested a method with a similar degree of the case structure and a history of predicate-argument structure analysis. He showed that the precision was improved compared them to a more conventional method. However, this method cannot be resolved appropriately in the history when there is no antecedent.

3.2. Metonymy Expression Resolution

The metonymy is a kind of metaphor and is a phenomenon to express by replacing a certain thing with the different things associated with it. For example, in the expression "漱石を読む"(read Soseki), it is thought that "漱石"(Soseki) points at "漱石の小説"(novels by Soseki). In the expression of "電源を入れる"(turn on the power), it is thought that "電源"(the power) points at "電源のスイッチ"(the power switch). Metonymy expression is the same as the form to be modifier B is omitted in the modification of the noun "A no B" type. In Japanese, there are many of this type of ellipsis that is not metonymy. It also appears in the differences in Figures 1 and Figures 2. For example, in the first sentence, "パソコン"(the PC) points at "パソコンのドライブ"(the drive of the PC) and "DVD"(the DVD) points at "DVD のディスク"(the disk of the DVD) in the expression of "パソコンに DVD を入れても"(even if I put the DVD in the PC). It is also important to deal with a phenomenon of the metonymy in anaphora resolution.

There is a study by Kiyota et al.[6] on research dealing with analysis of metonymy expression. The metonymy expression is known to become a cause of failure of the dependency parsing.For example, "電源"(the power) relate to "入れる"(turn on) in "電源を入れる"(turn on the power) which is metonymic expression. On the other hand, in a metonymic interpretative expression "電源"(the power) relate to "スイッチ"(the switch) and "スイッチ"(switch) relate to "入れる"(turn on), in "電源のスイッチを入れる"(turn on the power switch). Since this modifier-head gap in this metonymic expression and metonymic interpretative expression will affect the matching between the query text and the target text in a text-based question answering system, Kiyota et al. proposed a method to cancel the modifier-head gap using metonymy-interpretation expression pairs that they extracted automatically. This method of Kiyota et al. was not purposed for solving ellipsis directly, but has been able to find a candidate antecedent beyond the context. However, they did not treat priority, namely, in which the candidate is more correct.

Table 2 showed the reappearance number in the context[5] of the argument extraction that we counted in the preliminary investigation and that we introduced in Chapter 2. This is obtained by averaging the number of appearances of the word the same as the argument of the case that we extracted surface case in the sentence of each data.

[5]We calculated one inquiry data as one context in "Yahoo! Chiebukuro data". And we calculated one paragraph as one context in "Wikipedia" data.

Table 2. Number of the reappearance in the context

Num. of cases	Above	Below	Whole
Yahoo!Chiebukuro (colloquial)	0.244	0.257	0.501
Wikipdia (non-colloquial)	0.400	0.410	0.810

From the Table 2, the reappearance rate of argument extraction in the whole context of "Yahoo! Chiebukuro data" was 50% and at most had a low value. This shows the difficulty of the resolution in the context of the colloquial inquiry texts are compared with "Wikipedia".

However, in order to deal with text that contains the colloquial inquiry text resolution of the case where the antecedent is not present anywhere in the overall context is required. It is usually difficult to resolve it.

4. Prediction of Ellipses

As described in Chapter 3, the conventional method cannot be performed when there is not an antecedent in the same context. Generally, it is necessary to expand the scope of the search beyond the context to find the omitted words that do not have a candidate within the context. However, expanding the scope of the search, the number of candidate words is greater and it is more difficult to select even if the correct candidate set of ellipsis word was previously given. Therefore, in this paper, under the situation that the candidate set was given, we examined the method by using topic modeling within language model to consider that the appearance probability of each word is not the same but changes depending on the context. Furthermore, we decided to use the extracting method of a metonymic expression pair that has been proposed in the study of the Kiyota et al.[6] that we introduced in Chapter 3 but we do not discuss how to collect candidate set in this paper.

In this chapter, we review "topic modeling"[7] and "latent Dirichlet allocation"(LDA)[8] which we used in an evaluation experiment as one of topic models. After that we introduce the method and the results of this evaluation experiment.

4.1. Topic Modeling

Recently, the method called "topic modeling" attracts more attention as one of the statistical modeling methods to acquire knowledge from large-scale and heterogeneous, a large quantity of text information. The basic concept of topic modeling is that the words of a document do not appear independently but appear based on latent topics.

Until now, as a language model that can deal with global context information, the "cash model" and the "trigger model"[9] were representative models. These modelings modeled the relations of a word and another word directly but topic modeling models the relations of a word and a latent topic hidden in the context. Topic modeling represents each document by the mixture distribution of multiple topics and each topic by the mixture distribution of multiple words. "Latent Dirichlet allocation"(LDA) is the representative topic model. In this paper, we suppose that there are multiple topics for one document and model it by using an LDA which acts as a kind of multi-topic model.

4.2. LDA; Latent Dirichlet Allocation

Blei[8] has proposed a technique LDA that introduced the Dirichlet prior distribution as the prior distribution of the multinomial distribution that represent a topic of a document. Recently, the usefulness of topic modeling has attracted attention, and LDA is known to work well. Based on the idea that a document is represented as random mixtures over latent topics where each topic is characterized by a distribution over words, LDA infer the probability distribution of the topic. Figure 3 shows the graphical model of LDA.

In the graphical model, random variables and parameters are represented vertex; their dependencies are represented by a directed edge.The shaded vertex indicates observed variables; the vertices of the other indicate latent parameters or latent variables. The number written down at the rectangle's corner indicates the repetition of the generation of the variable in the rectangle. D is the number of documents, K is the number of topics and N_d is the word count in document d. θ and ϕ is a multinomial distribution parameter of the topic and multinomial

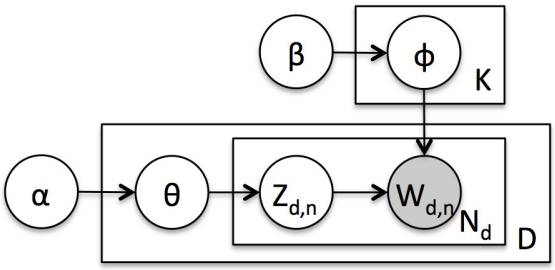

Fig. 3. Graphical model of LDA

distribution parameter of the word in each topic, respectively. α and β is Dirichlet hyperparameters in θ and ϕ, respectively. The document generative processes by this graphical model are as follows:

1. For each of the topics $k = 1, \cdots, K$:
   (a) Choose $\phi_k \sim Dir(\beta)$,
2. For each of documents $d = 1, \cdots, D$:
   (a) Choose $\theta_d \sim Dir(\alpha)$,
   (b) For each of the words $w_{d,n}$ *where* $n = 1, ..., N_d$:
      i. Choose a topic $z_{d,n} \sim Multi(\theta_d)$,
      ii. Choose a word $w_{d,n} \sim Multi(\phi_{z_{d,n}})$,

where ϕ_k is the word distribution for topic k, θ d is the topic distribution for document d, $z_{d,n}$ is the topic for the n-th word in document d, and $w_{d,n}$ is the n-th word in document d. $Dir(\cdot)$ is the Dirichlet distribution for parameter α, $Multi(\cdot)$ is the Multinomial distribution for parameter β. According to this LDA model, the total probability of the model is given:

$$P(\mathbf{W}, \mathbf{Z}, \theta, \phi|\alpha, \beta) = \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}}) \tag{1}$$

To infer the unknown parameters, there are various methods that have been proposed; using "Collapsed Gibbs sampling" of Griffiths et al.[10] are known to model estimation can be performed with high accuracy if the sufficient number of iterations has been obtained.Though there are θ and ϕ in the LDA model, Collapsed Gibbs sampler collapses (integrates out) these and derive the updating formula of the form that there are not θ and ϕ. The updating formula derived by Collapsed Gibbs sampling is given bellow:

$$P(z_i|z_{\backslash i}, \mathbf{W}) \propto \frac{p(w|z)p(z)}{p(w_{\backslash i}|z_{\backslash i})p(z_{\backslash i})} \frac{(n^v_{\backslash i,j} + \beta)(n^d_{\backslash i,j} + \alpha)}{(n^{\cdot}_{\backslash i,j} + \mathbf{W}\beta)(n^d_{\backslash i,\cdot} + \mathbf{T}\alpha)} \tag{2}$$

where $z_{\backslash i}$ indicates that it does not include the current assignment z_i from topic set Z. $n^v_{\backslash i,j}$ is the count of word v in topic j, that does not include the current assignment z_i, a missing subscript or superscript (e.g. $n^{(\cdot)}_{\backslash i,j}$) indicates a summation over that dimension. We can calculate a predictive distribution of topic distribution θ for each document and a predictive distribution of word distribution ϕ of each topic are calculated using samples obtained by Gibbs sampling. Equation 3 is estimated quantity $\hat{\theta}^k_d$ of the probability that topic k is generated in document d and Equation 4 is estimated quantity $\hat{\phi}^w_k$ of the probability that word w when topic k was chosen.

$$\hat{\theta}^k_d = \frac{N_{dk} + \alpha}{N_d + \alpha K} \tag{3}$$

$$\hat{\phi}^w_k = \frac{N_{kw} + \beta}{N_k + \beta V} \tag{4}$$

4.3. Choice from Candidate Words Using LDA

Using LDA, we introduce the idea that the appearance probability of each candidate word is not the same but changes depending on the context. Therefore, by the input of the number of topics K and sets of text (as global information) expressed by the sequence of the word $w(\in V)$, we can estimate the probability distribution $P(w|z_k)(w \in V)$ of the word w in each topic $z_k(k = 1, ..., K)$ and the probability distribution $P(z_k|d)$ of the topic $z_k(k = 1, ..., K)$ in each document d. In this report, by the input of the inquiry text d including an ellipsis, we examine four methods to choose a candidate word wi(wi $\in$ C,i,...I) from the candidate set C of the I.

1. Choice using the probability of the word (by N-gram)
2. Choice using the probability of the word (by LDA)
3. Choice using the probability of the word (by N-gram & LDA)
4. Choice using the classifier by the probability of the topic (by ML classifier)

Method 1 calculates appearance probability of candidate words w_i by the N-gram model[6] in an input inquiry text d and chooses a candidate word w_i of highest appearance probability. Method 2 calculates appearance probability of the candidate words w_i by the LDA model in an input inquiry text d and chooses a candidate word w_i of highest appearance probability. Method 3 calculates scores by linear interpolation of appearance probability of method 1 and method 2 described with Equation 5 and chooses a candidate word w_i of highest score.

$$score_{w_i}(d) = \lambda P_{ngram}(w_i|d) + (1 - \lambda)P_{lda}(w_i|d) \tag{5}$$

Method 4 uses machine learning (ML) classifier. In advance, we prepare a set of inquiry texts d' including the answer word w_i and each answer word $w_i(\in C)$ for training. Then the ML classifier learns the inquiry texts d' by the answer word w_i as the class. At this time, we use the mixing probability $P(z_k|d')$ of the topics in the text d' estimated by the LDA model as its feature. Then, we infer the mixed probability $P(z_k|d)$ of the topic in an input inquiry text d by that LDA model and choose the class w_i classifying it in a class by that ML classifier.

4.4. Experiments Using Artificial Data

In the evaluation experiment, we use the extracting method of a metonymic expression pair that has been proposed in the study of the Kiyota et al. that we introduced in Chapter 3. This is the method using pattern matching based on an appearance pattern of metonymy expression and the metonymy interpretation expression. Using this pattern, we extract all pairs of metonymic - interpretative expression from experimental data. Then, we collect the sets of the word that there are in both metonymic expression and metonymic interpretative expression in that as candidate sets C.

In this experiment, we find a pattern of metonymic interpretative expression in the test data and replace it with the metonymic expression of the pair beforehand. We predict the word that is missing by this substitution as a correct answer.

We have used 3,206,559 of the inquiry text data for "Yahoo! Chiebukuro data" collected in 4/1/2004~4/7/2009 for experimental data. Furthermore we have extracted 576,841 data of the subcategory "PC" from this, and have divided the 500,000 data by 100,000 data, equaling five. Then, we have evaluated it by a cross-validation using four of five for estimation of the LDA model and using one for evaluation. 3,8538 pairs of metonymy expression - interpretation expression[7] have been made from these 576,841 data. We have chosen two expressions as the target of the experiment. These expressions are the pairs that the total of the number of the appearances of the interpretation expression is higher, and ambiguity that these have is often confusing empirically.

We have shown the top 10 candidate words of expression 1 in Figure 4, the top 10 candidate words of expression 2 in Figure 5. Expression 1 has had a bias in the number of appearances of each candidate word, and expression 2 has tended to appear evenly candidate word.

[6] This is a language model typical approximated by N-th order Markov process the occurrence of the word.

[7] 38,538 pairs of metonymy expression - interpretation expression have been generated from 167,816 metonymy interpretation expression and metonymy expression 1,259,662.

Table 3. Expression for this experiment

No.	Expression	Appearances	Candidates	Sample of expression
1	“元の [candidate words] に戻す”	776	134	“元の状態に戻る” → “元に戻る”
2	“パソコンの [candidate words] について”	2,294	225	“パソコンの操作について” → “パソコンについて”

Table 4. Top 10 candidate words of expression 1

Candidate	Appearances	Ratio
状態	202	26.03%
サイズ	187	24.10%
画面	31	3.99%
位置	31	3.99%
色	31	3.99%
設定	26	3.35%
場所	23	2.96%
表示	18	2.32%
メモリ	12	1.55%
バージョン	12	1.55%

Table 5. Top 10 candidate words of expression 2

Candidate	Appearances	Ratio
メモリ	219	9.55%
購入	184	8.02%
ＣＰＵ	145	6.32%
電源	126	5.49%
スペック	117	5.10%
メモリー	86	3.75%
キーボード	84	3.66%
バッテリー	77	3.36%
画面	75	3.27%
こと	72	3.14%

In addition, we have used the implementation by “GibbsLDA++”[11] for the learning and the inference of the LDA model. We measured the perplexity of the probability distribution necessary to estimate the LDA model to determine the number of the topics K and hyper-parameters α, β of Dirichlet distribution using a part of the experimental data in advance. Then, we decided the number of topics $K = 150$ and hyper parameters $\alpha = 0.1, \beta = 0.01$ and $\beta = 0.01$ with the relevant perplexity of the model[8]. We have used the implementation by “SRILM tool kit”[12] for the N-gram model in method 2[9]. As for linear interpolation coefficient λ of Equation 5, we decided $\lambda = 0.85$ for expression 1 and $\lambda = 0.5$ for expression 2 that showed high accuracy by a prior experiment. We have used data mining software “WEKA”[13] in machine learning in method 3[10].

4.5. Results and Discussion

776 expression 1 of “元の [candidate words] に戻す”(go back up for [candidate words]) and 2,294 expression 2 of “パソコンの [candidate words] について”(about [candidate words] of the PC) have been found in the 500,000 experimental data. The accuracy rates that we calculated by the results that we predicted by four different methods have been shown in Tables 6 and 7 for each of two different expressions.

Table 6. Experimental result for expression 1

Type of method	N-gram	LDA	N-gram & LDA	ML classifier
Accuracy rate	38.02%	12.76%	40.59%	42.00%
Defference (based on N-gram)	-	-25.26	2.58	3.98

Table 7. Experimental result for expression 2

Type of method	N-gram	LDA	N-gram & LDA	ML classifier
Accuracy rate	11.38%	26.68%	33.35%	34.00%
Defference (based on N-gram)	-	15.30	21.97	22.62

From Table 6, method 2 (only LDA) has shown the accuracy rate that has been lower than method 1 (only N-gram). Method 3 (N-gram & LDA) has shown the accuracy rate that has been slightly higher than method 1 (only N-gram). Method 4 (ML classifier) has shown the highest accuracy rate. From Table 7, method 2 (only

[8] Using the one that was divided into five cross-validation to measure the perplexity of combination $K = \{30, 50, 150, 200\}$, $\alpha = \{0.01, 0.1, 0.5, 1, 1.5\}, \beta = \{0.01, 0.1, 0.5, 1, 1.5\}$

[9] We used 5-gram probability for the N-gram model, and use “interpolate” for the complementary model and “kndiscount” for the smoothing method as a parameter.

[10] We chose the LogitBoost that showed the highest classification performance from comparison with prior investigation, and used the default values for other options.

LDA) has shown the accuracy rate that has been higher than method 1 (only N-gram). Method 3 (N-gram & LDA) has shown a higher accuracy rate. Method 4 (ML classifier) has shown the highest accuracy rate again.

For method 1, compared to Table 6 and Table 7, expression 2 (has tended to appear evenly candidate word) has shown the accuracy rate that has been lower than expression 1 (has had a bias in the number of appearances of each candidate word). On the other hand, for method 2, expression 2 (has tended to appear evenly candidate word) has shown the accuracy rate that has been higher than expression 1 (has had a bias in the number of appearances of each candidate word). This is considered that the LDA model works better on the choice of candidate words than the N-gram model when there is no bias in the number of occurrences of each candidate word.

For method 2, the accuracy rate has been low in Table 6 and Table 7. It is thought that this does not consider an appearance position of the ellipsis in the test data when we calculate the appearance probability of the candidate word for test data. For method 3, it is thought that this considers an appearance position by N-gram model and it is effective to resolve the ambiguity of the word choice candidate that cannot be resolved in the N-gram model by LDA partially. For method 4, it is thought that this modifies the mismatch of topics information and candidate words by learning the mixing probability of topic information to each candidate word, and there is effective to strongly tie a topic to candidate word. But, it is thought that this is not a robust method because this does not consider an appearance position.

From these results, it has been confirmed that LDA model has the effect of improving the performance of the candidate word choice by complementing N-gram model that is a conventional language model and the effect of improving the performance of the selection of candidate words by tying candidate words to topics information using the machine learning.

5. Conclusion

In this study, we have aimed to assist in the accurate understanding of intentions in inquiry text collected via web forms and e-mail. And we have examined the method to predict the ellipsis that has focused on omitted word that is a feature of the colloquial style of Japanese text. We have also shown the results of evaluation experiments.

From the results of evaluation experiments, under the situation where the candidate set was given, we have confirmed that some of our methods using the LDA model have predicted the omitted word in the accuracy rate of more than 40% in the experimental results we have shown. Finally, the improvement of prediction accuracy, and the support of other ellipsis expression are future works.

Acknowledgements

We would like to thank Yahoo Japan Corporation and The National Institute of Informatics who provide us the data set of Yahoo! Chiebukuro.

References

[1] D. Kawahara, S. Kurohashi, Case Frame Construction by Coupling the Predicate and its Closest Case Component, *Journal of Natural Language Processing*, 2002, 9 (1), 3–19
[2] T. Masuoka, Y. Takubo, Kiso nihongo bunpoo [Basic Japanese grammar], *Tokyo: Kurosio*, 1992.
[3] B. J. Grosz, Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics* 21 (2) ,1995, 203–225.
[4] M. Walker, M. Iida, S. Cote, Japanese Discourse and the Process of Centering, *Computational Linguistics* 20 (2), 1996, 38.
[5] H. Yuta, K. Mamoru, M. Yuji, Improving japanese inter-sentential predicate argument structure analysis with contextual information and similarity between case structures, *IPSJ SIG Notes* 2011 (10), 2011, 1–8.
[6] Y. Kiyota, S. Kurohashi, F. Kido, Resolution of modifier-head relation gaps using automatically extracted metonymic expressions, *Journal of natural language processing* 11 (4), 2004, 127–145.
[7] T. Hofmann, Probabilistic latent semantic indexing. *SIG-IR'99*, 1999, 50–57.
[8] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *The Journal of Machine Learning* 3, 2003, 993–1022.
[9] Jelinek, Frederick, Statistical methods for speech recognition, *Bradford Books* 1997.
[10] Griffiths, T L, Gibbs sampling in the generative model of latent dirichlet allocation, *Unpublished note http://citeseerx ist psu edu/ . . .*, 2002.
[11] Xuan-Hieu Phan, Cam-Tu Nguyen, GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA), 2007
[12] Stolcke, A, SRILM-an extensible language modeling toolkit, *. . . conference on spoken language processing* 2002
[13] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, *Kaufman Series in Data Management*, Morgan Kaufmann, 2005.