An Empirical Study for Generating Zero Pronoun in Korean based on Cost-based Centering Model

Ji-Eun Roh
Div. of Electrical
and Computer Engineering
Pohang University of Science and Technology and Advanced Information Technology Research Center (AITrc)
San 31, Hyoja-dong, Nam-gu, Pohang, 790-784, R. of KOREA
jeroh@postech.ac.kr
fax: +82-54-279-5699

Jong-Hyeok Lee
Div. of Electrical
and Computer Engineering
Pohang University of Science and Technology and Advanced Information Technology Research Center (AITrc)
San 31, Hyoja-dong, Nam-gu, Pohang, 790-784, R. of KOREA
jhlee@postech.ac.kr
fax: +82-54-279-5699

Abstract

In Korean, in order to generate a coherent text, a redundantly prominent noun should be replaced by a non-zero pronoun or zero pronoun. Otherwise, the text becomes unnatural. Specifically, a redundant noun in Korean is frequently omitted while a redundant noun in English is replaced by a pronoun. This paper proposes a generation algorithm of the zero pronoun, using a *Cost-based Centering Model* which considers the inference cost. For an objective evaluation of our algorithm, we collected 87 texts from three genres, and manually recovered the omitted elements. Using the collected texts, we verify that our algorithm is well defined to explain the phenomenon of the zero pronoun in Korean. We also show that the proposed approach resolves both the over-generation of the zero pronoun in *Continue* and its under-generation in other transitions in terms of Centering.

1 Introduction

Text generation is the process of producing comprehensible texts in natural languages from non-linguistic representation of information. To generate a coherent text, we must pay attention to each stage of generation, such as content determination, text structuring, aggregation, and generation of anaphoric expression (pronominalization). Among these stages, we are especially interested in the generation of anaphoric expression focusing on zero pronouns, because generating the appropriate zero pronoun is directly connected with text coherence. Consider the following short text.

(1) **Na**-nun (I, topic) cinyel-toyn (on display) **os** (the dresses) cwung (one of) maum-ye tu-nun kes-i iss-ess-ta (attracted me).
(One of the dresses on display attracted me.)

(2) **[Na**-nun (I, topic), ø**]**[1] **[os**-oul (dress, object), ø**]** ip-e po-ass-ta (putted on).
(I (ø) putted it (ø) on.)

(3) Haciman (however), **[na**-nun (I, topic), ø**]** **[os**-i (dress, subject), ø**]** nemu (too) khesu (big) **[os**-ul (it, object), ø**]** sal-swu (buy) eps-ess-ta.(can not).
(However, the dress (ø) was too big so that I (ø) cannot buy it (ø))

In the above text, 'na (I)' appears repeatedly as a topic[2] in sentence (1), (2), and (3), and 'os (dress)'

[1] A bracketed noun, which means the unexpressed argument of the verb, is a *zero pronoun*. Generally, this kind of omitted element caused by the *zero anaphora* phenomenon is called *zero pronoun*, *zero element, zero anaphor*, or *null element*. In this paper, we call the omitted element a *zero pronoun.*

appears repeatedly as an object and a subject in sentence (2) and (3). Korean is a highly context-dependent language, and any arguments recoverable from the context are freely dropped. In the above text, ellipsis of redundant nouns, 'na (I)' and 'os (dress)', in sentence (2) and (3) is recommended to generate a natural Korean text. Otherwise, the text is not coherent because of redundancy.

Our goal is to generate natural anaphoric expressions in Korean, particularly the zero pronoun, using a *Cost-based Centering Model* which considers the inference cost. In this paper, the cost-based centering model refers to the revised centering model by Strube and Hahn (1999), which extends the original 4 transition types to 6 types and defines the cost between transition pairs with respect to the cost for inferring.

This paper is organized as follows. In Section 2, we describe the centering model, which is the main background knowledge for our algorithm to generate anaphoric expression. In Section 3, we briefly describe related works on the generation of anaphoric expressions. In Section 4, we investigate the characteristics of the zero pronoun in Korean, and in Section 5 we describe a cost-based centering model and our proposed algorithm. In Section 6, we discuss the experimental validation. Finally, in Section 7, we summarize the features of our work and future work.

2 The Centering Model

The centering model (Grosz et al., 1986; 1995) provides a framework for the interaction of cohesion and salience in the internal organization of a text. The model is formalized in terms of Cb, the *backward-looking center*, Cf, a list of *forward-looking centers* for each utterance U_n, (i.e., n^{th} utterance or sentence), and Cp, *preferred center* which is the most salient candidate for subsequent utterances. $Cf(U_n)$, i.e., the entities mentioned in U_n are ranked by some measures such as a grammatical role. $Cp(U_n)$ is the highest ranked center of $Cf(U_n)$ and is predicted to be $Cb(U_{n+1})$. If two successive utterances have no reference in common, the second will have no Cb.

Transition type across pairs of adjacent utterances is defined in terms of two factors: cohesion and salience. Cohesion is achieved if the $Cb(U_{n-1})$ and the $Cb(U_n)$ are the same, and salience is achieved if the $Cb(U_n)$ and the $Cp(U_n)$ are the same.

The model consists of three constraints and two rules.

Constraints

1. There is precisely one Cb in U_n.
2. Every element of $Cf(U_n)$ must be realized in U_n.
3. $Cb(U_n)$ is the highest-ranked element of $Cf(U_{n-1})$ that is realized in U_n.

Rules

1. If some elements of $Cf(U_{n-1})$ are realized as a pronoun in U_n then so is $Cb(U_n)$.
2. Transition types are ordered. *Continue* is preferred over *Retain,* which is preferred over *Smooth-Shift,* which is preferred over *Rough-Shift*.

	$Cb(U_n)=Cb(U_{n-1})$ or undefined $Cb(U_{n-1})$	$Cb(U_n)!=Cb(U_{n-1})$
$Cb(U_n)=Cp(U_n)$	Continue	Smooth-shift
$Cb(U_n)!=Cp(U_n)$	Retain	Rough-shift

Table 1. Transition Types

Although the centering model is attractive to NLP researchers, several issues remain. Several studies have been made on Cf-ranking (e.g., Strube and Hahn, 1999; Turan, 1998; Cote, 1998), because Cf-ranking is language-dependent. Cf-ranking for Korean is different from Cf-ranking for English in that two languages have different features in terms of word-order and functional typology. In this paper, we followed the Cf-ranking proposed by Roh (2003).

topic > subject > *directly-associated-entity (DAE)* > dir-obj > indir-obj > *immediately pre-verbal entity (IPV)*

The topic-first principle is attributed to the topic prominence of Korean.

3 Related Work

Several studies focus on the problem of anaphoric expression generation. The most primitive method

[2] Korean is a topic-prominent language. Topic, an element which is attached topic marker 'un/nun', not only marks the grammatical function of the head noun, but also adds some special meaning to them like "only", "also/too", "even", "in contrast to."

used for early anaphoric generation (e.g., McDonald, 1980; McKeown, 1985) is to use a simple rule: if the current sentence contains the same word mentioned in the previous sentence, use a pronoun to refer to the word. However, this simple rule tends to over-generate pronouns, which causes serious ambiguity.

Recently, some studies attempted to use the centering model for the generation of anaphoric expression. Kibble (2000) used the model to plan coherent texts and to select anaphoric expressions. He considered different strategies for choosing when to use a pronoun, and found the following to be the best: pronominalize the Cb only after a Continue. However, he did not provide experimental results to verify that the method is superior to other strategies.

Mitsuko et al. (2001) adopted the centering model to generate the zero pronoun in Japanese. In English, all arguments of a verb must be expressed in a sentence, and redundant arguments used in previous sentence are usually replaced by pronouns. However, Japanese allows arguments to be freely omitted when they are recoverable from a given context. Korean is quite similar to Japanese from this perspective. Mitsuko argued that all Cb are generated as zero pronouns in either Continue or Smooth-Shift transitions. This can be interpreted as they prefer the zero pronoun when salience rather than cohesion obtains. However, they did not explain the reason why they regarded salience rather than cohesion as an important factor in the zero pronoun.

4 Zero Pronoun in Korean

From the perspective of interpretation of zero pronoun, several studies (e.g., Kim, 1994; Kim, 1999; Ryu, 2000) about zero pronoun were performed by linguists in Korea. Most Korean linguists agree that the zero pronoun generally comes from the continuity of topic, salience of topic, and redundancy of discourse.

More concretely, from the perspective of information structure[3] proposed by Vallduvi (1990), Kim (1999) investigated the conditions of the zero pronoun in Korean: redundant *focus*, redundant *link*, and redundant *tail*. However, these conditions cannot be applied easily, because the focus, link, tail, and their redundancy are not automatically detected, and are determined pragmatically rather than structurally.

Kim (1999) also proposed certain conditions when the zero pronoun would be prohibited, claiming that old information that changes its role is not omitted. For example, when old information in the current sentence becomes a new topic in the next sentence, the role of the old information changes from the old focus to link for the new topic. This situation frequently occurs in the process of topic transition to expand the content of text. From the perspective of Centering, this condition can be interpreted to mean that the Cp of Retain which was one element of the previous sentence cannot be omitted in the transition sequence of Continue, Retain, and Smooth-Shift. Because this transition sequence, called a *Topic-Shift-Sequence* in this paper, is used to smoothly change the current topic to a new topic, and the new topic is generally realized as Cp in Retain. This will be examined as one of the hypotheses to generate anaphoric expression in the next Section.

Ryu (2000) investigated the zero pronoun in terms of Centering. He postulated that the zero pronoun is used to continue the center in Korean, i.e., zero pronoun usually comes from the Cb of Continue. This is supported by many Korean linguists. He also clarified that the zero pronoun rarely appears in written texts when compared with spoken texts.

In Ryu's experimental results, the zero pronoun in Smooth-Shift is worthy of attention. He counted the zero pronouns which comes from the Cb of each transition in four kinds of texts—written, quasi-written, quasi-spoken, and spoken, respectively. In written texts, only 6% of Cb in Smooth-Shift is omitted, and 86% is overtly expressed as a topic with topic marker. Similarly, in quasi-written texts only 12% of Cb in Smooth-Shift is omitted. This phenomenon contrasted with Mitsuko 's generation policy of zero pronoun: generate Cb as a zero pronoun in Smooth-Shift. Perhaps this differ-

[3] In information structure, the sentence is articulated into a trinomial hierarchical structure consisting of 'focus' and 'ground', with the latter further subdivided into 'link' and 'tail'. The focus corresponds to new information unknown to the hearer within a sentence. The ground is the complement of the focus. A link is an address pointer in the sense that it directs the hearer's knowledge-store, which is the information-anchoring role of the ground. The tail is the complement of the link within the ground.

ence is caused by the characteristics of the texts used in the experiments.

To summarize previous research related to zero pronoun in Korean, we conclude that the zero pronoun generally comes from the Cb of Continue. However, some problems still remain from the generation perspective.

- ♦ Which of Cb in Continue is omitted or not, among Cb of Continues?
- ♦ Ellipsis of Cb in the other transitions except for Continue
- ♦ Pronoun generation except for zero pronoun of Cb and Cf

According to Ryu's experimental results, only 26% of Cb in Continues is omitted from written texts. This means that only a partial portion of Cb in Continues becomes a zero pronoun. Recall that all previous research which used the centering model to generate pronouns or zero pronouns follow this principle: pronominalize (or omit in Japanese) Cb in all Continues. Considering Ryu's experimental results and our experimental results (see Section 6 for more details), this traditional strategy causes a serious over-generation of pronouns, including zero pronouns.

Concerning the second problem, almost all previous research considers only pronominalization in Continue, except for Mitsuko (2000) who included Smooth-Shift as well as Continue. However, we confirm that the zero pronoun of Cb in the other transitions also occurs from our corpus test. Therefore, the previous strategy causes the under-generation of pronouns including zero pronouns in the other transitions except for Continue.

Concerning the final problem, in Korean the redundant noun is generally realized as the original one rather than as a pronoun when the zero pronoun is forbidden, unlike English. Consider the following English sentence.

> I love my mother and I cannot imagine the world without **her**.

(1) Na-nun (I) wuli (my) emeni-lul (mother) salang-hako (love) **kunye (her)** eps-nun (without) sesang-un (the world) sangsang-hal-su eps-ta (cannot imagine).

(2) Na-nun (I) wuli (my) emeni-lul (mother) salang-hako (love) **emeni (mother)** eps-nun (without) sesang-un (the world) sangsang-hal-su eps-ta (cannot imagine).

In a English-to-Korean translation, sentence (2) which translates 'her' into 'emeni (mother)' is more natural than sentence (1) which translates 'her' into 'kunye (her)'[4]. In this paper, we consider only two types of noun expression, the original noun and the zero pronoun because of these kinds of Korean-dependent characteristics.

5 Generation of Zero Pronoun

5.1 Cost-based Centering Model

We propose a generation algorithm for anaphoric expression in Korean, particularly a zero pronoun, using a cost-based centering model. The model is the revised centering model by Strube and Hahn (1999), which extends the original 4 transition types to 6 types and defines the cost between transition pairs, with respect to the cost for inferring.

	$Cb(U_n)=Cb(U_{n-1})$ or undefined $Cb(U_{n-1})$	$Cb(U_n)!=Cb(U_{n-1})$
$Cb(U_n)=Cp(U_n)$ and $Cb(U_n)=Cp(U_{n-1})$	Cheap-Continue[5] (CC)	Cheap-Smooth-Shift (CSS)
$Cb(U_n)=Cp(U_n)$ and $Cb(U_n)!=Cp(U_{n-1})$	Expensive-Continue (EC)	Expensive-Smooth-Shift (ESS)
$Cb(U_n)!=Cp(U_n)$	Retain (R)	Rough-Shift (RS)

Table 2. Revised Transition Types (Strube, 1999)

Strube and Hahn (1999) argue that a Smooth-Shift which comes from $Cb(U_n) \neq Cp(U_{n-1})$ is less smooth, i.e., the Smooth-Shift requires a high processing cost, because it contradicts the intuition that a Smooth-Shift fulfills the prediction of the Retain. The same applies to a Continue with this characteristic. For this reason, they separated Expensive-Continue and Expensive-Smooth-Shift from Continue and Smooth-Shift in accordance with the equality of $Cb(U_n)$ and $Cp(U_{n-1})$, as shown in Table 2. These postulations coincide with our intuition.

[4] Kunye (her), corresponding to 'her' in English, is the third personal pronoun referring to a woman in Korean.

[5] In this paper, *Continue* and *Smooth-Shift* among revised transition types in Strube's work (1999) are called *Cheap-Continue* and *Cheap-Smooth-Shift* to distinguish from *Expensive-Continue* and *Expensive-Smooth-Shift*.

In this paper, in order to handle the zero pronoun, six transition types which are shown Table 2 and '*resume*' proposed by Knott et al. (2001) were applied. When an utterance mentions an entity not in the immediate previous utterance, but in the previous discourse, a *resume* occurs.

Many researchers working on the centering model agree that considering adjacent transition pairs rather than particular transition provides a more reliable picture about coherence and anaphora resolution (e.g., Grosz et al., 1995; Strube and Hahn, 1999; Kibble and Power, 1999; 2000). More concretely, Strube and Hahn (1999) proposed to classify all the occurrences of transition pairs with respect to the implied inference costs.[6] In this paper, the pair whose cost is cheap is called a *Preferred Transition Pair*, such as (CC,CC), (CC,R), (R,CSS), etc.

Cheap	(CC,CC), (CC,R), (R,CSS), (CSS,CC), (RS,CSS)
Expensive	Other pairs

Table 3. Cost of Transition Pairs

We investigate some transition pairs proposed by Strube and Hahn (1999), and partially adopt them to generate anaphoric expression, as shown in Table 3.

5.2 Generation Algorithm of Zero Pronoun

As the first step in our proposal, we must examine the reason why the revised transition types and preferred transition pairs considering inference cost are appropriate to the generation of the zero pronoun. We argue that the Cb of Cheap-Continue is more redundantly prominent than the Cb of Expensive-Continue. This assumption is reasonable if we consider that Expensive-Continue follows Retain which smoothly changes the topic. The prominence of Cb in Expensive-Continue decreases because of Retain. The following text, extracted from our corpus, is a description of an exhibition 'Cakwi (a kind of Korean traditional farming tools)', and is a good example to illustrate this phenomenon.

(1) Cakwi-nun (Cakwi, topic) wuli-nala-uy (Korean) centhong-cekin (traditional) nong-kikwu-i-ta (farming tool is) (Cakwi is a Korean traditional farming tool.)

(2) Cakwi-uy (Cakwi, adnom) nal-un (edge, topic) celsaknal (celsaknal) ilako-hanta (is called). (Edge of Cakwi is called celsaknal.)

➔ **CP : nal (edge, topic) CB : Cakwi (adnominal), *R***

(3) Cakwi-nun (Cakwi, topic) hyengtay-ka (shape, subject) dokki-wa (axe) pisus-hata (is similar to). (Cakwi is similar to that of an axe.)

➔ **CP : Cakwi (topic) CB : Cakwi (topic), *EC***

(4) Cakwi-nun (Cakwi, topic) khuki-ey ttala (by its size) tay-cakwi (big-cakwi), socakwi-lo (small-cakwi) nanwin-ta (is categorized). (Cakwi is categorized as big-cakwi and small-cakwi by its size.)

➔ **CP : Cakwi (topic) CB : Cakwi (topic), *CC***

In the above text, the topic smoothly changes from 'Cakwi' to 'nal (edge of Cakwi)' in sentence (2), and Retain occurs. This implies that the topic of the next sentence is 'nal', and it decreases the prominence of Cb, 'Cakwi', in sentence (2). However, in sentence (3), the topic is returned to 'Cakwi' from 'nal', and Expensive-Continue occurs. In sentence (4), 'Cakwi' is maintained as topic, and Cheap-Continue occurs. In this situation, it is natural that the Cb of sentence (3), 'Cakwi', is less prominent than Cb of sentence (4), 'Cakwi', even though both 'Cakwi' are the same as Cb of Continue transitions. If 'Cakwi' in sentence (3) is omitted, the topic (or subject) of 'pisus-hata (is similar to)' can be misinterpreted as 'nal' not 'Cakwi'.

The basic idea of anaphor generation is that the more the noun is redundantly prominent, the more the noun is pronominalized (or omitted).[7] Accordingly, we postulate that the Cb of Expensive-Continue is less elliptical than that of Cheap-Continue. For this reason, we adopt revised transition types considering the inference cost in order to generate the zero pronoun. The case of Smooth-Shift can also be explained in the same manner. For the same reason, it is reasonable that Cb of Continue which follows Continue is more prominent than the Cb of Continue which follows Rough-Shift. For this reason, we adopt the concept of preferred transition pairs.

With these issues in mind, we first construct the following assumptions to generate anaphoric expressions.

[6] Strube and Hahn (1999) argue that, given a sequence of utterances, *inference cost* is needed to understand them. They claim that the inference cost is 'cheap' if successive facts realize preferred transition pairs; otherwise, it is 'expensive'.

[7] Generally, redundantly prominent noun corresponds to Cb within a sentence.

(1) Do generate zero pronoun minimally in written texts.
(2) Do not pronominalize for new information.
(3) Do not make a zero pronoun when it causes ambiguity.
(4) $Cb(U_n)$ is more elliptical than $Cf(U_n)$.

Based on the above assumptions, our algorithm to generate zero pronoun is as follows.

(1)If $tr(U_n)$ = CC and $cost(U_{n-1},U_n)$ = cheap **then**
realize $Cb(U_n)$ as zero pronoun
(2)Else if $tr(U_{n-1})$ = CC and $tr(U_n)$ = R and $tr(U_{n+1})$ = CSS (i.e., if three sentences belong to Topic-Shift-Sequence) **then**
do not realize $Cp(U_n)$, $Cb(U_n)$, and $Cb(U_{n+1})$ as zero pronoun
(3)Else if ($tr(U_n)$ = R or $tr(U_n)$ = CSS) and $cost(U_{n-1},U_n)$ = cheap **then**
realize $Cb(U_n)$ as zero pronoun
(4)Else
do not realize $Cb(U_n)$ as zero pronoun

$tr(U_n)$: center transition of n^{th} sentence
$cost(U_{n-1},U_n)$: cost of transition pairs between $tr(U_{n-1})$ and $tr(U_n)$

Figure 1. Generation Algorithm of Zero Pronoun

Compared with the traditional anaphor generation strategy related Continue, the rule (1) which adopts an inference cost is more restrictive.

The following example text, which describes an exhibition 'Paymili (a kind of Korean traditional timber tool)', is a good example to illustrate rule (2). In this text, the topic changes from 'Paymili' to 'Namaksin (wooden shoes)' using the Topic-Shift-Sequence. In this process of topic change, Cp of sentence (3) occurring Retain, 'Namaksin', should not be omitted in order to imply topic change, and Cb, 'Paymili', had better not be omitted in order to smoothly change an old topic (Cb in sentence (3)) to a new topic (Cp in sentence (3)). Similarly, Cb which is equal to Cp in sentence (4) occurring Cheap-Smooth-Shift, 'Namaksin', had better not be omitted in order to emphasize a new topic. Therefore, we argue Cp and Cb of Retain and Cb(=Cp) of Cheap-Smooth-Shift as an unadvisable zero pronoun condition under the Topic-Shift-Sequence.

(1) Paymili-nun (Paymili) wuli-nala-uy (Korean) centhong-cekin (traditional) mokcey (timber) yencang-i-ta (tool is). (Paymili is a Korean traditional timber tool.)
(2) [Paymili-nun (topic)**, ø]** namaksin-ul (wooden shoes) kkak-ul (cutting) ttay (when) naypu-uy (inside) hyengtay-lul (shape) cap-nuntey (when forming) ssu-in-ta (is used). (Paymili is used for forming the inside shape when cutting wooden shoes.)
➔ **CP : Paymili (topic) CB : Paymili (topic), *CC*, cost(1,2) : cheap**
(3) Namaksin-un (wooden shoes) **Paymili**-lo (with Paymili) sin-uy (shoe's) moyang-ul (shape) kolu-ko (raking and), Hopikhal-ul (Hopikhal) iyong-hay (using) kkakk-nun-ta (cutting). (Wooden shoes are made by raking in the shape of a shoe with Paymili and cutting using Hopikhal)
➔ **CP : Namaksin (wooden shoes, topic) CB : Paymili (adverb), *R*, cost(2,3) : cheap**
(4) Namaksin-un (wooden shoes) pi o-nun (rainy) nal (days) cwulo (usually) sin-ess-nun-tey (are worn and) otong-namwu-na (paulownia tree or) petunamu-lo (willow from) mantul-ess-ta (are made). (Wooden shoes are usually worn on rainy days and are made from the paulownia tree or willow).
➔ **CP : Namaksin (wooden shoes, topic) CB : Namaksin (wooden shoes, topic), *CSS*, cost(3,4) : cheap**

According to the experimental results of Ryu (2000), the ellipsis ratio of Cb is proportional to the order of Continue, Retain, Smooth-Shift, and Rough-Shift. However, the ellipsis ratio in Retain and Smooth-Shift are low compared with that in Continue, and there is only a slight difference between the ellipsis ratio of Retain and that of Smooth-Shift. Obviously, the ellipsis ratio in Rough-Shift is too low. Therefore, we exclude Rough-shift, and propose rule (3) for Retain and Smooth-Shift under the condition that they do not belong to the Topic-Shift-Sequence. In this algorithm, we do not consider the ellipsis of other $Cf(U_n)$ except for $Cb(U_n)$.

6 Experiments

For an objective evaluation of our proposed algorithm, we investigated the phenomenon of Cb ellipsis from real texts. We collected 87 texts with 15 sentences on average, from three genres, news, story, and descriptive texts. The descriptive texts were gathered from the on-line museum site, 'the national folk museum of Korea'.[8] We manually recovered the omitted elements of collected texts. In this process, we did not recover the generic pronoun 'wuli (we)'.

As shown in Table 4, without the inference cost, 175 out of 374 Cb in Continues are realized as zero

[8] Our proposed algorithm will be used to upgrade the XExplainer system (Roh, 2001) which produces a *description* for commodities in Korean. The collected texts are similar to the domain of XExplainer in that they *describe* each exhibition. For this reason, we choose descriptive texts.

pronoun, i.e., 46% of Cb in Continues is omitted. With inference cost, the number, 374, is in turn divided into two classes: 203 Cheap-Continues and 171 Expensive-Continues, and 151 Cb out of 203 Cheap-Continues are omitted. Therefore, 86% (151/175) zero pronouns come from Cb of Cheap-Continues, not of Expensive-Continues. However, by considering 25% ((203-151)/203) of Cb in Cheap-Continues are not omitted, the issue remains concerning which Cb of Cheap-Continue should be omitted or not among the set of Cheap-Continues.

Transition	Without inference cost	With inference cost
CC	175(374)[9], 46%	151(203), 74%
EC		24(171), 14%
R	59(218), 27%	
CSS	34(86), 39%	29(47), 61%
ESS		5(39), 12%
RS	10(104), 9%	

Table 4. Ellipsis of Cb in each Transition

Transition pairs (X : any transition)		Cost of transition pairs	
		Cheap	Expensive
CC	(X, **CC**)	144(172), 83% ➔ **(Rule 1)**	7(31), 22%
R	(X, **R,** ¬CSS)	21(44), 47% ➔ **(Rule 3)**	36(162), 22%
	(CC, **R**, CSS)	2(12), 16% ➔ **(Rule 2)**	
CSS	(¬R, **CSS**)	27(35), 77% ➔ **(Rule 3)**	
	(CC, R, **CSS**)	2(12), 16% ➔ **(Rule 2)**	

Table 5. Ellipsis of Cb in Transition Pairs

The answer can be found in Table 5. Here, 172 out of 203 Cheap-Continues belong to cheap pairs, and the remaining 31 belong to expensive pairs. Moreover, there are 144 ellipses of Cb out of 172 Cheap-Continues associated with cheap pairs, and there are only 7 ellipses of Cb out of 31 Cheap-Continues associated with expensive pairs. To summarize, 82% (144/175) of zero pronouns in Continues come from the Cheap-Continues associated with cheap pairs. Accordingly, we roughly estimate that Cb in Cheap-Continue associated with cheap pairs is realized as a zero pronoun. Although this conclusion causes a slight under-generation of zero pronoun from the viewpoint of our experimental results, this satisfies our first assumption. If we follow the traditional anaphor generation strategy, generating Cb as a zero pronoun in all Continues, 374 zero pronouns occur from Continues. This causes an excessive over-generation of the zero pronoun. However, our algorithm generates only 172 zero pronouns from Continues. Accordingly, we conclude that rule (1) in Figure 1 is a good indicator for the ellipsis of Cb in Continues, and is more restrictive and elaborate than traditional strategies.

Concerning Retain, 27% Cb of total Retains is omitted, and more concretely, 47% Cb of Retains, which are associated with cheap pairs and which do not belong to the Topic-Shift-Sequence, is omitted. 88% Cb of Retains, which are associated with cheap pairs and which belong to the Topic-Shift-Sequence, is not omitted

Concerning Smooth-Shift, without the inference cost, 39% Cb of total Smooth-Shifts is omitted. With inference cost, the number, 86, is in turn divided into two classes: 47 Cheap-Smooth-Shifts and 39 Expensive-Smooth-Shifts. Considering the ellipsis ratio, Cb in Cheap-Smooth-Shift is more elliptical than the Cb in Expensive-Smooth-Shift. More concretely, 77% Cb of Cheap-Smooth-Shifts, which are associated with cheap pairs and which do not belong to the Topic-Shift-Sequence, is omitted. 88% Cb of Cheap-Smooth-Shifts, which are associated with cheap pairs and which belong to the Topic-Shift-Sequence, is not omitted..

Therefore, we estimate that Cb of Retain and Cb of Cheap-Smooth-Shift in the Topic-Shift-Sequence are not realized as zero pronouns, as indicated rule (2) in Figure 1. Additionally, we found that 92% Cp of Retains which belong to Topic-Shift-Sequence is not omitted.

However, rule (3) in Figure 1 is open to discussion because compared with our experimental results, it causes the over-generation of zero pronoun from Cb in Retains. However, in the case of Smooth-Shift, it is effective. Compared with approach of Mitsuko (2001), rule (3), generation of zero pronoun in Smooth-Shift is more elaborate without excessive over-generation.

[9] Parenthesized number means the total number of transitions, and the number in front of the parenthesis means the number of transitions in which Cb is omitted. The percent means the ratio of the two numbers.

7 Conclusion

In this paper, we propose an algorithm for the generation of anaphoric expression, the zero pronoun, in Korean. Our algorithm is based on the cost-based centering model, which extends transition types and defines the cost of transition pairs with respect to the inference cost. Using the model, we resolve both the over-generation of the zero pronoun in Continue and its under-generation in other transitions. We also propose a rule in which ellipsis of Cb or Cp is inadvisable. According to our experimental results, the Cb of cheap transition is more elliptical than that of expensive transition. In addition, the Cb of transition associated with cheap pairs is more elliptical than that of a transition associated with expensive pairs.

This paper did not handle the ellipsis of Cf elements except for Cb. The Pronoun Rule of the centering model applies only to the anaphoric expression which is the Cb of the current sentence. With respect to all other anaphoric expressions except for Cb in the current sentence, the centering model is under-specified. However, there are many omitted elements which are not Cb in our experiment, and they are left as future work. Additionally, the practicality of the proposed method will also be verified through a more reliable evaluation methodology in a real generation system.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc), and also partially by the Brain Korea 21 Project in 2003.

References

Cote, S. 1998. *Ranking Forward-Looking Centers*, Centering Theory in Discourse. Oxford: Clarendon Press, pp55-71

Grosz, B.J. and Sidner, C.L. 1986. *Attention, intentions, and the structure of discourse,* Computational Linguistics, pp175-203

Grosz, B.J., Joshi, A.K. and Weinstein, S. 1995. *Centering: A Framework for Modeling the Local Coherence of Discourse*, Computational Linguistics 21(2), pp203-225

Kibble, R. and Power, R. 1999. *Using centering theory to plan coherent texts*, In Proceedings of the 12th Amsterdam Colloquium.

Kibble, R. and Power, R. 2000. *An integrated framework for text planning and pronominalisation* Proceedings of the 1st International Conference on Natural Language Generation (INLG-2000), Mitzpe Ramon, Israel, pp77-84

Kim, M. K., 1999. *Conditions on Deletion in Korean based on Information Packaging*, Discourse and Cognition 1(2), pp61-88

Kim, M.Y. 1994. *The Centering of Korean discourse*, Ms thesis, Seoul National University

Knott, A., Oberlander, J., O'Donnell, M. and Mellish, C. 2001. *Beyond elaboration: the interaction of relations and focus in coherent text*, In T. Sanders, J. Schilperoord and W. Spooren (eds.) Text representation: linguistic and psycholinguistic aspect. Amsterdam: Benjamins, pp181-196

McDonald, D.D. 1980. *Natural Language Production as a Process of Decision Making under Constraint*, Ph.D. thesis, MIT.

McKeown, K.R. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge, U.K.: Cambridge University Press.

Mitsuko,Yamura-Takei, Fujiwara, M., and Aizawa, T. 2001. *Centering as an Anaphora Generation Algorithm: A Language Learning Aid Perspective*, NLPRS 2001, Tokyo, Japan, pp557-562

Roh, J.E., Kang, S.J. and Lee, J.H. 2001. *Korean Text Generation from Database for Homeshopping Sites*, NLPRS 2001, Tokyo, Japan, pp419-426

Roh, J.E. and Lee, J.H. 2003. *Coherent Text Generation using Entity-based Coherence Measures*, ICCPOL, Shen-Yang, China, pp243-249

Ryu, Byung Ryul, 2001, *Centering and Zero Anaphora in the Korean Discourse,* Seoul National University, Ms Thesis

Strube, M. and Hahn, U. 1999. *Functional Centering: Grounding Referential Coherence in Information Structure* Computational Linguistics 25(3), pp309-344

Turan, Umit D. 1998. *Ranking Forward-Looking Centers in Turkish: Universal and Language Specific Properties,* Centering Theory in Discourse. Oxford: Clarendon Press, pp139-160

Vallduvi, E. 1990. *The Informational component, doctorial dissertation*, University of Pennsylvania