Automatic Voice Selection in Japanese based on Various Linguistic Information

Ryu Iida and **Takenobu Tokunaga**
Department of Computer Science, Tokyo Institute of Technology
W8-73, 2-12-1 Ookayama Meguro Tokyo, 152-8552 Japan
{ryu-i,take}@cl.cs.titech.ac.jp

Abstract

This paper focuses on a subtask of natural language generation (NLG), *voice selection*, which decides whether a clause is realised in the active or passive voice according to its contextual information. Automatic voice selection is essential for realising more sophisticated MT and summarisation systems, because it impacts the readability of generated texts. However, to the best of our knowledge, the NLG community has been less concerned with explicit voice selection. In this paper, we propose an automatic voice selection model based on various linguistic information, ranging from lexical to discourse information. Our empirical evaluation using a manually annotated corpus in Japanese demonstrates that the proposed model achieved 0.758 in F-score, outperforming the two baseline models.

1 Introduction

Generating a readable text is the primary goal in natural language generation (NLG). To realise such text, we need to arrange discourse entities (e.g. NPs) in appropriate positions in a sentence according to their discourse salience. Consider the two following Japanese texts, each of which consists of two sentences.

(1) *Tom$_i$-wa kouen$_j$-ni it-ta .*
Tom$_i$-TOP park$_j$-IOBJ go-PAST
(Tom$_i$ went to a park$_j$.)

Kare$_i$-wa soko$_j$-de ookina inu-ni oikake-rareta .
he$_i$-TOP there$_j$-LOC big dog-IOBJ chase-PASSIVE/PAST
(He$_i$ was chased by a big dog there$_j$.)

(2) *Tom$_i$-wa kouen$_j$-ni it-ta .*
Tom$_i$-TOP park$_j$-IOBJ go-PAST
(Tom$_i$ went to a park$_j$.)

Ookina inu-ga soko$_j$-de kare$_i$-o oikake-ta .
big dog-SUBJ there$_j$-LOC he$_i$-OBJ chase-PAST
(A big dog chased him$_i$ there$_j$.)

In (1), 'Tom$_i$' is topicalised in the first sentence, and then it appears at the subject position in the second sentence. In contrast, the same argument, i.e. 'he$_i$' is realised at the object position in the second sentence of text (2). Intuitively, text (1) is relatively more natural than text (2). Thus, given the two predicate argument relations, go(SUBJ:Tom$_i$, IOBJ:park$_j$) and chase(SUBJ:big dog, OBJ:Tom$_i$, IOBJ:park$_j$), a generation system should choose text (1).

The realisation from a semantic representation (e.g. predicate argument structures) to an actual text has been mainly developed in the area of natural language generation (Reiter and Dale, 2000), and has been applied to various NLP applications such as multi-document summarisation (Radev and McKeown, 1998) and tutoring systems (Di Eugenio et al., 2005). During the course of a text generation process, various kinds of decisions should be made, including decisions on textual content, clustering the content of each clause, discourse structure of the clauses, lexical choices, types of referring expressions and syntactic structures. Since these different kinds of decisions are interrelated to each other, it is not a trivial problem to find an optimal order among these decisions. This issue has been much discussed in terms of architecture of generation systems. Although a variety of architectures has been proposed in the past, e.g. an integrated architecture (Appelt, 1985) and a revision-based architecture (Inui et al., 1994; Robin, 1994), a pipeline architecture is considered as a consensus architecture in which decisions are made in a predetermined order (Reiter, 1994). Voice selection is a syntactic decision that tends to be made in a later stage of the pipeline architecture, even though it influences various decisions, such as discourse structure and lexical choice. Unlike referring expression generation, voice selection has received less attention and been less discussed in the past. Against this background, this

Proceedings of the 14th European Workshop on Natural Language Generation, pages 147–151,
Sofia, Bulgaria, August 8-9 2013.

research tackles the problem of voice selection considering a wide range of linguistic information that is assumed to be already decided in the preceding stages of a generation process.

The paper is organised as follows. We first overview the related work in Section 2, and then propose a voice selection model based on the four kinds of information that impact voice selection in Section 3. Section 4 then demonstrates the results of empirical evaluation using the NAIST Text Corpus (Iida et al., 2007) as training and evaluation data sets. Finally, Section 5 concludes and discusses our future directions.

2 Related work

The task of automatic voice selection has been mainly developed in the NLG community. However, it has attracted less attention compared with other major NLG problems, such as generating referring expressions. There is less work focusing singly on voice selection, but not entirely without exception, such as Abb et al. (1993). In their work, passivisation is performed by taking into account both linguistic and extra-linguistic information. The linguistic information explains passivisation in an incremental generation process; realising the most salient discourse entity in short term memory as a subject eventually leads to passivisation. In contrast, extra-linguistic information is used to move a less salient entity to a subject position when an explicit agent is missing in the text. Although these two kinds of information seem adequate for explaining passivisation, their applicability was not examined in empirical evaluations.

Sheikha and Inkpen (2011) focused attention on voice selection in the generation task distinguishing formal and informal sentences. In their work, passivisation is considered as a rhetorical technique for conveying formal intentions. However, they did not discuss passivisation in terms of discourse coherence.

3 Voice selection model

We recast the voice selection task into a binary classification problem, i.e. given a predicate with its arguments and its preceding context, we classify the predicate into either an active or passive class, taking into account predicate argument relations and the preceding context of the predicate.

As shown in examples (1) and (2) in Section 1, several factors have an impact on voice selection in a text. In this work, we take into account the following four information as features The details of the feature set are shown in Table 1.

Passivisation preference of each verb An important factor of voice selection is the preference for how frequently a verb is used in passive sentences. This means each verb has a potential tendency of being used in passive sentences in a domain. For example, the verb '*yosou-suru* (to expect)' tends to be realised in the passive in the newspaper domain because Japanese journalists tend to write their opinions objectively by omitting the agent role. To take into account this preference of verb passivisation, we define a preference score by the following formula:

$$score_{pas}(v_i) = \frac{freq_{pas}(v_i)}{freq_{all}(v_i)} \cdot \log freq_{all}(v_i) \quad (1)$$

where v_i is a verb in question[1], $freq_{all}(v_i)$ is the frequency of v_i appearing in corpora, and $freq_{pas}(v_i)$ is the frequency of v_i with the passive marker, *(ra)reru*. The logarithm of $freq_{all}(v_i)$ is multiplied due to avoiding the overestimation of the score for less frequent instances. In the evaluation, the preference score was calculated based on the frequency of each verb in the 12 years worth of newspaper articles, which had been morpho-syntactically analysed by a Japanese morphological analyser Mecab[3] and a dependency parser *CaboCha*[4].

Syntactic decisions As described in Section 1, various kinds of decisions are interrelated to voice selection. Particularly, syntactic decisions including voice selection directly impact sentence structure. Therefore, we introduce syntactic information except for voice selection which prescribes how an input predicate-argument structure will be realised in an actual text.

Semantic category of arguments Animacy of the arguments of a predicate has an impact on their syntactic positions. Unlike in English, inanimate subjects tend to be avoided in Japanese. In order to capture this tendency, we use the semantic category of the arguments of the verb in question (e.g.

[1]Note that the preference needs to be defined for each word sense. However, we here ignore the difference of senses because selecting a correct verb sense for a given context is still difficult.

[1]*Bunsetsu* is a basic unit in Japanese, consisting of at least one content word and more than zero functional words.

[2]http://nlp.cs.nyu.edu/irex/index-e.html

[3]https://code.google.com/p/mecab/

[4]https://code.google.com/p/cabocha/

type	feature	definition
PRED	score_{pas}	passivisation preference score defined in equation (1).
	lexical	lemma of P.
	func	lemma of functional words following P, excluding passive markers.
SYN	sent_end	1 if P appears in the last *bunsetsu*[1]-unit in a sentence; otherwise 0.
	adnom	1 if P appears in an adnominal clause; otherwise 0.
	first_sent (last_sent)	1 if P appears in the first (last) sentence of a text; otherwise 0.
	subj(obj,iobj)_embedded	1 if the head of the adnominal clause including P is semantic subject (object, indirect object) of P; otherwise 0.
ARG	subj(obj,iobj)_ne	named entity class (based on IREX[2]) of the subject (object, indirect object) of P.
	subj(obj,iobj)_sem	semantic class of the subject (object, indirect object) of P in terms of Japanese ontology, *nihongo goi taikei* (Ikehara et al., 1997).
COREF	subj(obj,iobj)_exo	1 if the subject (object, indirect object) of P is unrealised and it is annotated as exophoric; otherwise 0.
	subj(obj,iobj)_srl_order	order of the subject (object, indirect object) of P in the SRL.
	subj(obj,iobj)_srl_rank	rank of the subject (object, indirect object) of P in the SRL.
	subj(obj,iobj)_coref_num	number of discourse entities in the coreference chain including P's subject (object, indirect object) in the preceding context.

P stands for the predicate in question. The four feature types (PRED, SYN, ARG and COREF) correspond to each information described in Section 3.

Table 1: Feature set for voice selection

named entity labels provided by *CaboCha*, such as Person and Organisation, and the ontological information defined in a Japanese ontology, *nihongo goi taikei* (Ikehara et al., 1997)) as features.

Coreference and anaphora of arguments As discussed in discourse theories such as Centering Theory (Grosz et al., 1995), arguments which have been already most salient in the preceding context tend to be placed at the beginning of a sentence for reducing the cognitive cost of reading, as argued in Functional Grammar (Halliday and Matthiessen, 2004). In order to consider the characteristic, we employ an extension of Centering Theory (Grosz et al., 1995), proposed by Nariyama (2002) for implementing the COREF type features in Table 1. She proposed a generalised version of the forward looking-center list, called the *Salient Reference List* (SRL), which stores all salient discourse entities (e.g. NP) in the preceding contexts in the order of their saliency. A highly ranked argument's entity in the SRL tends to be placed in the subject position, resulting in a passive sentence if that salient entity has a THEME role in the predicate-argument structure. To capture this characteristic, the order and rank of discourse entities in the SRL are used as features[5].

In addition, as described in Abb et al. (1993), if the agent filler of a predicate is underspecified, the passive voice is preferred so as to unfocus the underspecified agent. Likewise, if the argument (in this case, the agent filler) of a predicate is exophoric, the passive voice is selected.

4 Experiments

We conducted an empirical evaluation using manually annotated newspaper articles in Japanese. To estimate the feature weights of each classifier, we used MEGAM[6], an implementation of the Maximum Entropy model, with default parameter settings. We also used SVM[7] with a polynomial kernel for explicitly handling the dependency of the proposed features.

4.1 Data and baseline models

For training and evaluation, we used the NAIST Text Corpus (Iida et al., 2007). Because the corpus contains manually annotated predicate argument relations and coreference relations, we used those for the inputs of voice selection. In our problem setting, we conducted an intrinsic evaluation; given manually annotated predicate argument relations and coreference relations of arguments, a model determines whether a predicate in question is actually realised in the *passive* or *active* voice in the original text. The performance is measured based on recall, precision and F-score of correctly detecting passive voice. For evaluation, we divided the texts in the corpus into two sets; one is used for training and the other for evaluation. The details of this division are shown in Table 2.

We employed two baseline models for compar-

[5]In Table 1 "*_srl_rank" stands for how highly the argument's referent ranked out of the discourse entities in the SRL, while "*_srl_order" stands for which slot (e.g. TOP slot or SUBJ slot, etc.) stores the argument's referent.

[6]http://www.cs.utah.edu/~hal/megam/

[7]http://svmlight.joachims.org/

	#articles	#predicates	#passive predicates
training	1,753	65,592	4,974 (7.6%)
test	696	24,884	1,891 (7.6%)

Table 2: Data set division for evaluation

	R	P	F
$\theta = 0.1$	0.768	0.269	0.399
$\theta = 0.2$	0.573	0.357	**0.440**
$\theta = 0.3$	0.403	0.450	0.425
$\theta = 0.4$	0.293	0.512	0.373
$\theta = 0.5$	0.161	0.591	0.253
$\theta = 0.6$	0.091	0.692	0.162
$\theta = 0.7$	0.060	0.717	0.111
$\theta = 0.8$	0.030	0.851	0.058
$\theta = 0.9$	0.014	1.000	0.027

Table 3: Effect of threshold θ for score$_{pas}$

ison. One is based on the passivisation preference of each verb. The model uses only $score_{pas}(v_i)$ defined in equation (1), that is, it selects the *passive* voice if the score is more than the threshold parameter θ; otherwise, it selects the *active* voice. The other baseline model is based on the information that the existence of an exophoric subject results in selecting the passive voice. To capture this characteristic, the model classifies a verb in question as *passive* if the annotated subject is exophoric; otherwise, it selects the *active* voice.

4.2 Results

We first evaluated performance of the first baseline model with various θ. The results are shown in Table 3, demonstrating that the baseline achieved its best F-score when θ is 0.2. Therefore, we set the θ to 0.2 in the following comparison.

Table 4 shows the results of the baselines and proposed models. To investigate the impact of each feature type, we conducted feature ablation when using the maximum entropy model (ME:* in Table 4). Table 4 shows that the model using the feature type PRED achieves the best performance among the four models when using a single feature type. In addition, by adding feature type(s), the F-score monotonically improves. Finally, the results of the model using the PRED, ARG and COREF features achieved the best F-score, 0.605, out of the two baselines and models based on the maximum entropy model. It indicates that each of the features except SYN feature contributes to improving performance in a complementary manner.

Furthermore, the results of the model using SVM with the second degree polynomial kernel show better performance than any model based on the maximum entropy model. This means that the combination of features is important in this task because of the dependency among the four kinds of information introduced in Section 3.

model	R	P	F
baseline1: score$_{pas} \geq 0.2$	0.573	0.357	0.440
baseline2: exophora	0.493	0.329	0.395
ME: PRED	0.270	9.612	0.374
ME: SYN	0.000	N/A	N/A
ME: ARG	0.095	0.516	0.161
ME: COREF	0.092	0.574	0.159
ME: PRED+SYN	0.282	0.618	0.387
ME: PRED+ARG	0.380	0.647	0.479
ME: PRED+COREF	0.480	0.762	0.589
ME: SYN+ARG	0.133	0.558	0.215
ME: SYN+COREF	0.147	9.618	9.238
ME: ARG+COREF	0.267	0.661	0.380
ME: PRED+SYN+ARG	0.397	0.656	0.494
ME: PRED+SYN+COREF	0.485	0.760	0.592
ME: PRED+ARG+COREF	0.506	0.752	**0.605**
ME: SYN+ARG+COREF	0.281	0.673	0.397
ME: ALL	0.507	0.747	0.604
SVM(linear): ALL	0.456	0.792	0.579
SVM(poly-2d): ALL	0.679	0.858	**0.758**

Table 4: Results of automatic voice selection

5 Conclusion

This paper focused on the task of automatic voice selection in text generation, taking into account four kinds of linguistic information: passivisation preference of verbs, syntactic decisions, semantic category of the arguments of a predicate, and coreference or anaphoric relations of the arguments. For empirical evaluation of voice selection in Japanese, we used the predicate argument relations and coreference relations annotated in the NAIST Text Corpus (Iida et al., 2007). Integrating the four kinds of linguistic information into a machine learning-based approach contributed to improving F-score by about 0.3, compared to the best baseline model, which utilises only the passivisation preference. Finally, we achieved 0.758 in F-score by combining features using SVM.

As future work, we are planning to incorporate the proposed voice selection model into natural language generation models for more sophisticated text generation. In particular, generating referring expressions and voice selection are closely related because both tasks utilise similar linguistic information (e.g. salience and semantic information of arguments) for generation. Therefore, our next challenge is to solve problems about generating referring expressions and voice selection simultaneously by using optimisation techniques.

References

B. Abb, M. Herweg, and K. Lebeth. 1993. The incremental generation of passive sentences. In *Proceedings of the 6th EACL*, pages 3–11.

Douglas E. Appelt. 1985. Planning English referring expressions. *Artificial Intelligence*, 26(1):1–33.

Barbara Di Eugenio, Davide Fossati, Dan Yu, Susan Haller, and Michael Glass. 2005. Natural language generation for intelligent tutoring systems: A case study. In *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pages 217–224.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

M. A. K. Halliday and C. Matthiessen. 2004. *An Introduction to Functional Grammar*. Routledge.

R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceeding of the ACL Workshop 'Linguistic Annotation Workshop'*, pages 132–139.

S. Ikehara, M. Miyazaki, S. Shirai A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei (in Japanese)*. Iwanami Shoten.

Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1994. Text revision: A model and its implementation. In *Aspects of Automated Natural Language Generation: Proceedings of the 6th International Natural Language Generation Workshop*, pages 215–230.

S. Nariyama. 2002. Grammar for ellipsis resolution in Japanese. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 135–145.

D. R. Radev and K. R. McKeown. 1998. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 24(3):469–500.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Ehud Reiter. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170.

Jacques Robin. 1994. *Revision-based Generation of Natural Language Summaries Providing Historical Background – Corpus-based Analysis, Design, Implementation and Evaluation*. Ph.D. thesis, Columbia University.

F. Abu Sheikha and D. Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193.