Extraction of Cognition Results of Travel Routes with a Thesaurus

Kazutaka TAKAO
Department of Global Development Science,
Graduate School of Science and Technology,
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe
657-8501, Japan
003d912n@y02.kobe-u.ac.jp

Yasuo ASAKURA
Dept. of Science of Regional and Built Env.,
Graduate School of Science and Technology,
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe
657-8501, Japan
asakura@kobe-u.ac.jp

Abstract

We are attempting to model travel route choice behaviour with language to describe the thinking process of travelers because words can directly and clearly reflect their psychological states from a bottom-up viewpoint. This paper shows a method that extracts impressions and feelings, i.e., cognition results of travel routes, out of open-ended questionnaire texts with a thesaurus. Complex words are also allowed as cognition results. Additional considerations and training contents are also reported. Finally, an experiment on the extraction of cognition results from unseen texts is reported.

1 Introduction

1.1 Background

We are attempting to linguistically model spatial cognition and travel route choice behaviour. In the field of travel engineering, a unit of travel called a "trip" is expressed as movement from an origin to a destination; route choice behaviour is expressed as choice from the set of alternatives in each trip.

Existing city and traffic infrastructure planning are mainly framed according to economical effects and demand estimations. Many studies have approached route choice behaviour from such top-down standpoints. These studies are more interested in the results of travel behaviour rather than the psychological state. Many have used numerical equations to explain behaviour quantitatively, and psychological factors are often expressed with internal variables.

On the other hand, when developing new products, a company must analyze customers' awareness from a bottom-up viewpoint. In the same way, it is also important for traffic infrastructure planning to analyze from a bottom-up viewpoint, observing travelers' psychological states when making choices. We expect to handle psychological factors more clearly with words. Psychological states of route choice behaviour can be explained by words with analyzing open-ended questionnaire texts.

1.2 Modeling of Travel Route Choice Behaviour

Route choice behaviour "recognizes" the characteristics of the alternatives (i.e. travel routes) in a choice set and "chooses" one route. Therefore, route choice behaviour can be expressed as a two-stage process that places the "cognition result" between the travel route and the behaviour as shown in Figure 1.

The first stage recognizes the characteristics, feelings, and impressions of each route in the

choice set, i.e., cognition results. The second stage chooses a route by evaluating cognition results. These stages are analyzed with open-ended questionnaire texts.

This paper describes the first stage, i.e., the "recognize" stage whose objective is to extract enough cognition results to handle the travel route choosing process in the second stage. This paper introduces a process that uses a thesaurus to extract cognition results from open-ended questionnaire texts.

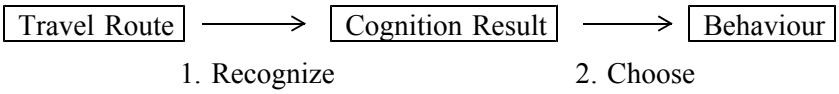

Figure 1: Route Choice Process

1.3 Aspect in EBA Model

Extracted cognition results are handled in the second stage. Therefore, we have to see the second stage to clarify the requirements of the first stage. The second stage is expressed by such decision-making models as Tversky's Elimination-By-Aspects (EBA) model (1972). In our study an "aspect" is a feature of a situation that shares several alternatives, such as "bright" and "comfortable." The cognition results correspond to aspects. In the EBA model, a decision is made by eliminating alternatives and judging whether each alternative includes the aspect in question. For example, when a traveler wants to choose a "comfortable" route, the routes that do not have the cognition result "comfortable" are eliminated from the choice set. If more than one route remains, the traveler considers another aspect in the next priority order. In this way, a final route is chosen by eliminating routes according to the priority order of aspects. The structure of decision making can be plainly analyzed from open-ended questionnaire texts because the words that represent such aspects as "comfortable" appear in them.

Consequently, the requirements of this paper are as follows. First, cognition results should be extracted as expressions of aspects. Therefore, we have to handle the meanings of words by considering what the questionnaire texts want to say rather than simply handling the words grammatically. Second, cognition results should be categorized easily because aspects indicate the categories of the meaning of words rather than external appearance.

Note that the word "aspect" in this paper reflects the meaning used in the EBA model rather than the normal grammatical meaning.

1.4 Related Studies

Some studies have attempted to analyze human psychological states from open-ended questionnaire texts. Inui (2004) analyzed open-ended questionnaire texts about traffic infrastructure planning, focusing on the intention behind answers. This paper, however, concentrates on analyzing how people feel.

Tateishi et al. (2002) and Kobayashi et al. (2004) extracted evaluations of specific products from the large quantity of language resources that exist on the Web. They easily found language resources using product names as keywords. However, evaluations of travel behaviour are difficult to collect from the Web because the awareness of habitual activity tends to be latent, and moreover, it is difficult to distinguish whether it expresses travel behaviour because it is lacking specific keywords. Hence, large language resources do not always exist in this research. Moreover, evaluations of aspects are analyzed from texts as priority orders rather than their method that used a

previously made dictionary.

2 Extraction Process

Using a thesaurus to extract cognition results from open-ended questionnaire texts is convenient for the following reasons. First, we can use the wealth of vocabulary covered in the thesaurus even if it does not appear in training texts. This is advantageous for obtaining high vocabulary coverage without collecting large amounts of training texts. Second, extracted cognition results can easily correspond to categories of the thesaurus to be classified into aspects.

We suggested a framework of extracting cognition results from open-ended texts with a thesaurus (See Takao and Asakura (2004-a) or (2004-d) for details). After determining the relationship between the thesaurus categories and cognition results, we can extract cognition results from unseen texts using the relationship as a template. The method of extraction is as follows:

- Step 1: Invest the semantic code of each morpheme by looking up the thesaurus.
- Step 2: Extract cognition results using the relationship.

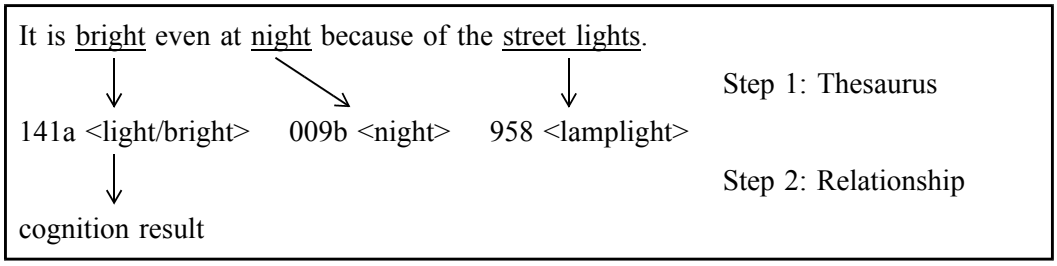

Figure 2: Extraction with a Thesaurus

We used the Kadokawa New Thesaurus released by Oono and Hamanishi (1989) as a thesaurus, Chasen as a morphological analyzer, and CaboCha as a dependency structure analyzer. The classification of a Kadokawa thesaurus is indicated with semantic codes. An example is shown in Figure 2. The semantic code of "bright" is 141a. If we know the relationship that code 141a expresses cognition results, we can judge "bright" as a cognition result. However, the following problems were found by simply applying the above steps:

(a) One morpheme is sometimes too short to understand as a cognition result, and hence thin meaning words were extracted. For example, the cognition result "time" can be extracted from the sentence "Waiting takes time." However, "time" is too short to be understood as an aspect.

(b) Incorrect cognition results were extracted from some expressions. For example, "sit" was extracted as a cognition result from the sentence: "I feel anxious about traffic jams even if I can sit." But this is wrong. The words in the expression "even if" should be ignored.

Therefore, the following considerations were added to the steps. First, we allowed complex words that consist of multi-morphemes as cognition results. Second, a judgment step of the expression patterns is added as follows:

- Step 3: Judge the expression patterns to filter out incorrect cognition results.

Cognition results are clarified by the allowance of complex words. On the other hand, the following demerits are anticipated:

(a) Complex words are usually not covered in a thesaurus. Hence, we have to collect the vocabularies ourselves.

(b) The standard length of morphemes becomes ambiguous. Hence, we allowed the following complex words as cognition results: words whose synonyms or antonyms are covered in the thesaurus.

Example 1 Synonyms (Japanese words are shown in italics):
Covered: *kokochiyoi* (comfortable).
Allowed: *kimochi* (feeling) *ga* (particle) *yoi* (good).
Example 2 Antonyms:
Covered: *teiji* (scheduled time).
Allowed: *touchaku-jikan* (arrival time) *ga* (particle) *fuantei* (unsettled).

Thus, we have to train the following contents:

(a) Vocabulary knowledge for Step 1.
(b) Relationship between the semantic codes and cognition results for Step 2.
(c) Expression patterns to avoid extracting incorrect cognition words for Step 3.

3 Data Set Preparation

We need to collect linguistic resources that include feelings and impressions. Newspaper texts are widely used in natural language processing (NLP), however, they are inappropriate for this study because the articles are usually written objectively without subjective views. Therefore, we must collect the data set ourselves.

Moreover, subjects are required to write actively what they feel about travel routes. Many existing studies that analyze travel behaviour have used observation equipments to collect numerical data. In such cases, the data is collected automatically even if the subjects are passive. On the other hand, subjects are required to be active in our study. Therefore, we must realize that the amount of linguistic resources in our case is not as large as other NLP tasks, such as machine translations, because of the difficulty of collection.

We conducted a questionnaire about going from a certain place to Kyoto City Hall (See Takao and Asakura (2004-b) for details). Four routes were shown as the choice set; bicycle, subway, bus, and taxi. Scenarios were presented in which such spatial conditions as season, weather, and time were different. The subjects were asked to write freely what they thought about each route and each scenario in open-ended texts. This is a questionnaire about transport mode choice in the narrow meaning. However, it can be treated within the travel route choice in the wide meaning because its process is also expressed as "recognize" and "choose."

1209 valid sentences are collected. 200 sentences were separated at random as the test set for unseen texts. 1009 sentences were used for training.

4 Trained Contents

4.1 Vocabulary

4.1.1 Limitation of Semantic Code(s)

Limiting the semantic codes of wide meaning words into a travel route choice task is important to filter out extra noise. For example, "*kakeru*" has meanings that include "hang," "lack," "run," "spend" etc. If this word is used in the expression "*kaban* (baggage) *wo* (particle) *kakeru*," the semantic code should be limited to "hang"; otherwise, such an inaccurate cognition result as

"spend" would be extracted.

4.1.2 Uncovered Words

We collected the vocabularies of uncovered words from training texts whose semantic codes were also trained by hand. This work is necessary for handling travel behaviour because some basic words are not covered in the thesaurus.

Examples:

gotoobi (every 5th and 10th day), *doatsuudoa* (door-to-door), *byuun* (mimetic word; fast), *akusesu* (access), *iratsuku* (annoying)

4.1.3 Complex Words

A collection of complex words means the combination of nouns, verbs, particles, etc. Therefore, many expressions have the same meaning. For example, the following words that mean "obviousness of travel time" are collected from the training texts as shown in Table 1.

Table 1: Collected Words that Mean "Obviousness of Travel Time"

jikan (time) *ga* (particle) *hakareru* (can estimate)
jikan (time) *ga* (particle) *yomeru* (readable)
jikan (time) *doori* (exactly)
shoyou-jikan (required time) *ga* (particle) *yomeru* (can be read)
ryokou-jikan (travel time) *no* (particle) *yosou* (expectation) *ga* (particle) *tate* (plan) *yasui* (with ease)
shoyou-jikan (required time) *no* (particle) *mikomi* (expectation) *ga* (particle) *tate* (plan) *yasui* (easy)
machi-jikan (waiting time) *ga* (particle) *wakaru* (know)
touchaku-jikan (arrival time) *ga* (particle) *keisan* (calculate) *dekiru* (can)

Sometimes, cognition results do not appear directly. In everyday conversations, feeling or impression words are often hidden by common sense and tacit understandings. In this paper, indirect expressions are treated as cognition results by interpreting them with common sense into direct expressions covered in the thesaurus and estimating those semantic codes. Some examples are shown in Table 2.

Table 2: Interpretation and Semantic Codes of Indirect Words

Indirect words	Interpretation	Semantic Code
nimotsu (baggage) *ga* (particle) *heru* (decrease)	*migaru* (agile)	693 <easy>
kyori (distance) *ga* (particle) *nagai* (long)	*tooi* (far)	108a <far>

4.2 Relationship

The Kadokawa thesaurus consists of 2814 categories. 109 categories were judged as the categories of cognition results. Table 3 shows the 10 topmost categories that appeared frequently in the training texts. Their total number of appearance and examples are also shown.

Table 3: 10 Topmost Categories of Cognition Results

Category	# of words	Example of Words
691a <unpleasant>	59	*taihen* (hard), *mendou* (troublesome), *uttoushii* (annoying)
171a <price>	58	*takai* (expensive), *hiyou ga kakaru* (cost), *yasui* (cheap)
155b <slow>	56	*osoi* (slow), *jikan ga kakaru* (time-consuming)
252a <wet>	55	*nureru* (get wet)
693 <easy>	42	*raku* (easy), *rakuchin* (easy)
166b <distinct>	42	*kakujitsu* (certain), *seikaku* (accurate)
691 <pleasant>	41	*kaiteki* (comfortable), *kimochi ga yoi* (comfortable)
146a <hot>	38	*atsui* (hot), *mushiatsui* (muggy), *atatakai* (warm)
146b <cold>	37	*samui* (cold), *tsumetai* (cold), *suzushii* (cool)
156c <early & late>	36	*hayai* (early), *osoi* (late)

4.3 Expression Patterns

Knowledge of special expressions helps avoid extracting incorrect cognition results. For example,

Chikatetsu no hou ga hayakat ta to koukai suru. (The subway would have been faster.)

In this sentence, "*hayakat* (fast)" expresses another route, not the focused route. Therefore, the words in the expression "*no hou ga ... ta*" should be ignored.

Another example follows:

Kakujitsu ni hayai. (It is certainly fast.)

"*Kakujitsu* (certainly)" is not a cognition result, i.e., certainty is not recognized. Rather, "*hayai* (fast)" is recognized, and "*kakujitsu*" only modifies "*hayai*." Therefore, modifiers that express the degree of cognition results should be ignored. We collected 14 expression patterns from the training texts.

5 Experiment on Unseen Texts

We carried out an experiment that extracted cognition results from unseen texts. 117 sentences different from training texts were selected out of the separated 200 sentences. The overlap ratio of the same sentences is high because some sentences consisted of only one word, for example "hot," and because some cognition results are common regardless of scenario. Only different sentences were chosen for this experiment to see the availability of extraction from absolutely unseen texts. The results are as follows:

Table 4: Experiment Results

	(A) Extracted correctly	(B) Partly extracted	(C) Extra noise	(M) Missing
Total	119	9	6	9
Insufficient training of...				
limitation of semantic codes	-	-	1	-
uncovered words	-	-	-	1
complex words	-	9	2	7
expression patterns	-	-	3	-
Morphological analyzer error	-	0	0	1

Table 4 shows the number of extracted or missing cognition results. The causes of the faults are also shown. If we treat (B) as faults, recall is 86.9% and precision is 88.8%.

Only one fault is caused by insufficient training of uncovered words. Many of the unseen simple words are covered in the thesaurus. Therefore, even if they do not appear in the training texts, cognition results have been correctly extracted by referring to the thesaurus. This means that the proposed method of using the thesaurus is effective. On the other hand, many of the faults are caused by insufficient training of complex words because these vocabularies are collected only by training texts. Finally, note that no faults are caused by insufficient training of relationships.

6 Discussion

6.1 Complex Words Collection

The problem of complex words collection remains. Because there are various expressions of complex words, it is inefficient to collect them from training texts, i.e., to wait passively for their appearance in the corpus. Rather, it may be more efficient to let some subjects write complex words of the same meaning actively by showing keywords. Moreover, since this task is similar to collecting paraphrasing words, such ideas as Takao et al. (2002) are useful.

Furthermore, it will be efficient to train as mappings from combinations of semantic codes with particles to semantic codes of cognition results. For example, "*yomeru*" and "*wakaru*" in Table 1 have the same semantic code 413a <cognition>, hence they can be trained with code 413a rather than with each vocabulary.

6.2 Denial

The denial of cognition results should be extracted to be understood properly as an aspect. It can be expressed as a modifier of cognition result categories. In some cases, cognition results were denied directly; however, in other cases, the action verb is denied, and the cognition result modifies the verb. In both cases, the extraction should be in the form of the "denial of the cognition result." Details will be reported in another paper.

Example:

"It is not hot" ⟶ not hot ⟶ denial of 146a <hot>

"I cannot arrive there easily" ⟶ not easily ⟶ denial of 693 <easy>

6.3 Causality of Cognition Results

This paper shows the extraction of cognition results. In the same way, such static attributes of travel routes as facilities and the dynamic conditions of travel space such as weather, season, time, etc. can be also extracted. Thus, the causality of generating cognition results can be extracted as rules. For example, the following rule, extracted from the text shown in Figure 2, is useful for travel infrastructure planners to give hints for modifying travel spaces.

Street light & night ⟶ bright.

7 Conclusion

We extracted cognition results from open-ended questionnaire texts. The accomplishments of this paper are briefly summarized as follows:

(1) We extracted cognition results of travel routes from open-ended questionnaire texts with a thesaurus with high accuracy.

(2) Not only the covered words in the thesaurus, complex words are also allowed as cognition results.

(3) The collection of complex words was insufficient because they are not covered in existing thesauri.

The results of this paper were used to analyze the second stage, i.e., the "choose" stage. See Takao and Asakura (2004-c) for details.

Acknowledgements

Our thanks are extended to everyone at the Institute of System Science Research for their kind cooperation with the questionnaire survey.

References

CaboCha. http://chasen.org/~taku/software/cabocha/.

ChaSen. http://chasen.naist.jp/hiki/ChaSen/.

Inui, H. 2004. A study on extraction of intention of open-ended questionnaire texts and automatic classification - mainly about request intention -. Ph.D. Thesis, Kobe University, Japan (in Japanese).

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T. 2004. Collecting evaluative expressions for opinion extraction. in *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pp. 584-589.

Oono, S. and Hamanishi, M. 1989. *Kadokawa New Thesaurus CD-ROM Version*, Kadokawa Shoten Publishing / Fujitsu, Japan.

Takao, K. and Asakura, Y. 2004-a. Extraction of cognition rules of travel routes with a thesaurus. in *Proceedings of The Tenth Annual Meeting of The Association for Natural Language Processing*, pp. 103-106 (in Japanese).

Takao, K. and Asakura, Y. 2004-b. Data collection for modeling spatial cognition and route choice behaviour linguistically. in *Proceedings of the 24th Conference of Japan Society of Traffic Engineering* (to appear, in Japanese).

Takao, K. and Asakura, Y. 2004-c. Extraction of choice strategy of aspects among travel routes from open-ended texts. in *Proceedings of the 30th Conference of Infrastructure Planning*, CD-ROM (to appear, in Japanese).

Takao, K. and Asakura, Y. 2004-d. Catching and modeling spatial cognition and route choice behaviour linguistically. in *Proceedings of the 9th Conference of Hong Kong Society for Transportation Studies (9th HKSTS)* (to appear).

Takao, K., Imamura, K., and Kashioka, H. 2002. Comparing and extracting paraphrasing words with 2-way bilingual dictionaries. in *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*, pp.1016-1022.

Tateishi, K., Morinaga, S., Yamanishi, K., and Fukushima, S. 2002. Analyzing opinions on the web -- a framework for combining information extraction with text mining --. in *Proceedings of the 64th Annual Meeting of Information Processing Society of Japan* Vol. 3, pp. 19-20 (in Japanese).

Tversky, A. 1972. Elimination by aspects: a theory of choice. *Psychological Review* Vol. 79 No. 4, pp. 281-299.