The Valence Patterns of Japanese Verbs Extracted From The EDR Corpus

Takano Ogino [*1;2] Hitoshi Isahara [*2;3] Kazuhiro Kobayashi [*1]

*1 EDR (Japan Electronic Dictionary Research Institute, Ltd.)
*2 Kobe University
*3 CRL (Communications Research Laboratory)
EDR,Akishima Building,
4-34-7, Ikebukuro, Toshima-ku
Tokyo 171-0014, Japan
Ogino@edr.co.jp kobayasi@edr.co.jp isahara@crl.go.jp

Abstract

This paper describes research on particular verb valences obtained from actual linguistic data.

We created verb valence data using data from the EDR Co-occurrence Dictionary as our source. The EDR Co-occurrence Dictionary is coded with syntactic governing-dependent relation tags and semantic tags. The syntactic governing-dependent relations data in the EDR Co-occurrence Dictionary however, is expressed as individual constituent pairs. In this study, we grouped each of the governing-dependent relation pairs according to their verb concepts and then unified them into a number of combinations based on case.

After the data was automatically unified from the source data, we manually corrected mistaken governing-dependent relations, and also made changes to case where necessary. By following this procedure, we created basic valence data for each verb. Further, based on this valence data and the verb patterns created from it, we are currently looking into creating semantic groups for nouns on which semantic restrictions are imposed by the verb.

1. Introduction

In Japanese, case is not indicated by the word order in a sentence, but rather by case-marking postpositions ('*joshi*' in Japanese). Verbs determine what case postposition is used and also restricts the semantic group of the noun that precedes the postposition. This is generally referred to as 'selectional restriction'. By extracting and analyzing this type of information from actual linguistic data, valence patterns for individual verbs are created. Valence patterns for Japanese verbs are expressed as follows:

Verb pattern⇒[N1, N2,…Ni…]
Ni⇒ <semantic group name for noun> + case postposition

Verb valence patterns can be used to make selections in machine translation when
there are multiple correspondence words and governing-dependent relations from which to choose. This analysis makes use of the property of verb-dependent valence patterns, which places restrictions on case postpositions and on the meaning of the nouns that precede the case postpositions.

EDR has a large tagged corpus that includes syntactic governing-dependent relations, word meaning (concept) labels and semantic role indicators for predicates.

The data used in this study comes from the EDR Co-occurrence Dictionary, which contains approximately 9000 differing words of verb, 14,000 differing concepts and an approximate total of 800,000 verb samples.

Using this data as a base, we extracted the verb valence patterns that existed in the corpus.

Valence is generally provided based on syntactic case relations only (1) or with syntactic case relations and a semantic classification of words that can appear before the case postpositions (2). In this study however, we also include as part of the dataset semantic role information (3) for verb classification.

(1) [subject-'*ga*', object-'*wo*']
(2) [subject-'*ga*' (human), object-'*wo*' (food)]
(3) [subject-'*ga*' (human): agent,
object-'*wo*' (food): object]

This paper will describe the following points regarding the valence patterns extracted from the EDR Corpus.

1) Background Information on Creating Valence Patterns
2) Resources Used In Creating Valence Data
3) Procedure Used In Creating Valence Patterns from the EDR Corpus
4) Verb Classification Derived From Valence Data
5) Semantic Classification for Nouns in Valence Patterns

2. Background Information on Creating Valence Patterns

What case postposition a verb takes and what meaning a noun preceding the case postposition has differs from verb to verb. Also, even when the combinations of surface case are the same, there are semantic restrictions on the nouns that precede the case, which places further limitations of the words that can be used.

For example, in machine translation, given a multi-sense Japanese word, the appropriate translation in a target language will differ according to the meaning when the word is used. If the intended meaning of the word in the

sentence can be determined automatically, it is possible to select the appropriate corresponding word in the target language. Valence patterns for specific verbs can be used in the aforementioned case. For example, the Japanese word "切る (*'kiru'*) has various meanings as shown below. The appropriate English translation is given in brackets{ }. (Japanese words are expressed in italic or Japanese characters.)

Example. 1

切る (*'kiru'*)

1) Cut <concrete object> with <cutting tool>. {cut}
2) Complete <telephone call>. {hang up}
3) Stop moving <machine> etc. {switch ~ off}

If you restrict the noun that co-occurs with the verb, it is possible to create a distinction. In Example 1, this is the information for the semantic group of the items shown in brackets, <concrete object, telephone, machine, etc.>.

3. Resources Used In Creating Valence Data

As part of the development of the EDR Electronic Dictionary, we developed a corpus of approximately 200,000 sentences. The information includes syntactic governing-dependent relation tags, and for those items thought to have a governing-dependent relation semantic tags are also included.

We also made use of the EDR Co-occurrence Dictionary. The Co-occurrence Dictionary was developed from the EDR corpus. The EDR Co-occurrence Dictionary data is comprised of syntactic governing-dependent pairs. One constituent of the pair is the governing constituent (the verb) and the other is the constituent being governed (the noun).

Example 2.

Example of the verb 'eat' and the object case postposition '*wo*'

JCC0016556
Syntactic Information:
ice cream (noun)[3bff86] =>['*wo*']=> eat (verb)[3bc6f0]
Semantic Information:
eat (3bc6f0) =>[object]=> ice cream (3bff86)
Example:
Sentence Number / The <ice cream> was (eaten).
Sentence Number / We (eat) <ice cream>.

JCC0024319
Syntactic Information:
breakfast (noun)[3be060] =>['*wo*']=> eat (verb)[3bc6f0]
Semantic Information:
eat (3bc6f0) =>[object]=> breakfast(3be060)
Example:
Sentence Number / We (eat) <breakfast>.
Sentence Number / We don't (eat) <breakfast>.

JCC0024697
Syntactic Information:
breakfast (noun)[3be060] =>['*wo*']=> eat (verb)[3bc6f0]
Semantic Information:
eat (3bc6f0) =>[object]=> breakfast(3be060)
Example:
Sentence Number / Let's start (eating) <breakfast>.

JCC0036808
Syntactic Information:
aphides (noun)[0e3672] =>['*wo*']=> eat (verb)[3bc6f0]
Semantic Information:
eat (3bc6f0) =>[object]=> aphides (0e3672)
Example:
Sentence Number / It (eats) <aphides>.

4. Procedure Used In Creating Valence Patterns from the EDR Corpus

4.1. Creating Valence Data

As shown in Example 2, we extracted the relations from the Japanese Co-occurrence Dictionary for the verbs and associated nouns. Since Example 2 is a binary relation, we pick up only one case relation.

Example 3

gohan wo taberu (eat dinner)
Sample Sentence 1
banana wo taberu (eat a banana)
Sample Sentence 2
…
…
imouto ga taberu (younger sister eats)
Sample Sentence 1
kodomo ga taberu (a child eats)
Sample Sentence 2
…
…

Based on this data, we collected the records with the same sample sentence number in the Co-occurrence Dictionary, the same verb notation, and the same concept id number. By collecting these records, we obtain the following:

Example 4

imouto ga gohan wo taberu.
(younger sister eats dinner)
Sample Sentence 1
kodomo ga banana wo taberu.
(a child eats a banana.)
Sample Sentence 2

4.2. Transforming Case

In this study, valence data was created using the base form of the verb. In the actual sentences however, there are sentences in which the passive, causative or potential verb

forms were used. For these cases, the case relation that is used generally differs from the case relation when the base form is used.

4.2.1. Expressions with the Potential

For Japanese dependency markers, it is not apparent from the surface expression what the actual case is. For example, '*ha*' can be used instead of '*wo*' and '*ni*' (to) can be used for '*kara*' (from) etc. For instances where this occurred, the actual case must be determined by reading the sentence. The arrow symbol (⇒) indicates a change to the base form:

Example 5
banana [mo] taberareru.
⇒ *banana [wo] taberu koto ga dekiru.*
(We can eat a banana. / A banana is eatable.)

Example 6
miruku [nara] nomeru.
⇒ *miruku [wo] nomu koto ga dekiru.*
(I can drink milk. / Milk is drinkable.)

If you consider the relation of the verb 'eat' in the sentence '*banana wo taberu koto ga dekiru*', we can determine that the case postposition is '*wo*'. For cases in which the surface case is not in the base form, it is necessary to manually correct the case to the base form as illustrated here.

4.2.2. Expressions with the Passive

As shown in Example 7, for sentences in which the passive is used, the sentences were changed to active sentences to obtain the base form.

Example 7

watashi [ha] haha [ni] homerareta.
(I was praised by my mother.)
⇒ *haha [ga] watashi [wo] hometa.*
(My mother praised me.)

4.2.3. Problems in Original Data

The syntactic governing-dependent relations in the EDR Corpus were created to a certain degree automatically. Therefore, there are some cases in which the relations were extracted in error. In addition, when the verbs in the collocational data (extracted from the governing-dependent binary relations) are automatically sorted, there are cases where it is not possible to determine if a postposition is a case postposition or an adverbial ending. This resulted in postpositions that did not actually represent case being included among those instances in which postpositions were actually used as case postpositions.

The Japanese case postposition '*ni*' for example, is used to indicate "GOAL" as in '*basho ni* (to a place)' or '*aite ni* (to a person)'. '*ni*' is also used as a suffix ending for adverbs and adjectives to indicate a state as in '*shizuka ni* (quietly)' or '*dandan ni* (gradually)'. In the latter case, the postposition '*ni*' is not used as an actual case postposition. Since the distinction between these two different instances cannot be made automatically from the surface string, both of these cases were extracted. However, for the purposes of obtaining valence data, we only want cases where '*ni*' was used to indicate "GOAL". Therefore, we manually eliminated those cases where the second usage of '*ni*' had been included.

5. Verb Classification Derived From Valence Data

As shown in Example 2.1, the EDR co-occurrence relations used in the EDR Co-occurrence Dictionary provide semantic role indicators such as 'agent', 'object', 'source', 'tool' etc. for the words in a sentence.

Example 2.1
Syntactic Information:
ice cream (noun)[3bff86] =>['*wo'*]=> eat (verb)[3bc6f0]
Semantic Information:
eat (3bc6f0) =>[object]=> ice cream(3bff86)

The information for each of the words in the sentence 'I eat ice cream', is as follows:

Word : { I }
Semantic Classification : <human>
Semantic Role : [agent]
Word : {ice cream}
Semantic Classification : <food>
Semantic Role : [object]

In order to classify the verbs, we combined
(1) the surface level case portion
(Content of [ ] from Syntactic Information above) and
(2) the semantic roles
(Content of [ ] from Semantic Information above) for each verb.

We then expressed the verb showing multiple case:

⇒[agent] '*ga'* [object] '*wo'* '*taberu*'(eat)

We analyzed 1400 concepts. This is approximately one tenth of the total dataset.

Number of Concepts: 1400
Total Number of Patterns: 2300
Total Number of Different Patterns: 750

By arranging the differing patterns hierarchically as shown in Table 1, we get a semantic classification of the verbs.

Level Number		Deep Level Case (Relation Label and Case Postpositions)	EDR's 3rd and 4th Level Hierarchy	Words for Case Pattern
1	1	agent (*ga*)	Physical Activity	運動 (exercise), 泳ぐ (swim)
			Movement	家出 (leave home; run away from home),
			Physical Act	働く (work), 休む (rest)
			Emotional Activity	沈黙 (be silent), 騒ぐ (make noise; be excited)
81	1	agent (*ga*) place (*ni*)	Physical Activity	座る (sit; be seated),
81	2	agent (*ga*) place (*wo*)	Physical Activity / Movement	歩く (walk), 埋める (bury; fill), 歩き回る (walk about; pace)
82	1	agent(*ga*) place-from (*kara*)	Movement	引き揚げる (withdraw; move out of; leave), 離れる (move away from)
82	2	agent (*ga*) place-from (*kara*) place (*de*)	Movement	降りる (move off from)
82	3	agent (*ga*) place-from (*wo*)	Movement	出る (leave), 離れる (move away from)
82	4	agent(*ga*) place-from (*wo*) place (*de*)	Movement	降りる (alight from; get off),
82	5	agent (*ga*) place-from (*kara*) place-to (*ni*)	Movement	避難 (escape from; take refuge)
82	6	agent(*ga*) place-from (*kara*) place-to(*e*)	Movement	向かう (go towards)

Table 1: Hierarchy of Verb Patterns

Of the patterns described in Table 1, a count can be measured for patterns that are higher in the hierarchy. The result is shown in Table 2.

Case Pattern	Upper Classification	Count of Subordinate Patterns		Count of Concepts	
agent(*ga*)	Volitional Act	150	20.1%	280	12.0%
agent(*ga*)object(*wo*)	Volitional Act ; Object	246	32.9%	1122	48.2%
object(*ga*)	Phenomenon	167	22.4%	538	23.1%
causal_potency(*ga*)object(*wo*)	Phenomenon ; Act of involuntary body; Object	49	6.6 %	148	6.4 %
cause(*ga*) object(*wo*)	Phenomenon; Causal act; Object	10	1.3 %	32	1.4 %
experiencer(*ga*)	Thought, emotion, psychological action ;Experiencer	49	6.6 %	94	4.0 %
time(*ga*)	Passage of time	5	0.7 %	6	0.3 %
a-object(*ga*)	Attribute, state	69	9.2 %	103	4.4 %
other		2	0.3 %	3	0.1 %
Total		747	100.0 %	2326	100.0 %

Table 2: Semantic Classification and Verb Patterns

6. Semantic Classification for Nouns in Valence Patterns

6.1. Abstracting Semantics for Nouns From Valence Data

There are approximately 800,000 records that comprise the valence data. This data was created by extracting verbs from approximately 200,000 sentences in the corpus and collecting the 14,000 differing concepts that are associated with the verbs. This is a rather large dataset. In analyzing natural language automatically, even though example-based systems, fast searches and increased memory capacity are available, it is still difficult to process language given such a large dataset.

To address this, we decided to reduce the number of described patterns by creating semantic groups for the nouns. EDR created a concept classification (the EDR Concept Classification Dictionary). However, to meet our current objectives, this classification is not yet sufficient. Therefore, we decided to try and create a semantic marker classification from the actual data we extracted. For example, in Japanese the words おぶう ('*obuu*') and 背負う ('*seou*') both mean to carry something on the back. '*seou*' can be used to refer to carrying 'people' or 'things'. The word '*obuu*' however, does not co-occur with 'things'.

The concept for '*obuu*' is narrower than the concept '*seou*'. In addition, there is a strong semantic restriction on the noun that can co-occur with '*obuu*'.

Example 8

荷物を背中に背負う。
(put luggage on one's back)
*荷物を背中におぶう。
(carry luggage on one's back)
子供を背中に背負う
(put a baby on one's back)
子供を背中におぶう。
(carry a baby on one's back)

(* ill formed expression)

It is necessary to decide what semantic groups to create for the nouns that are restricted by verbs. Currently, we are sorting through the nouns from the valence data we obtained from the linguistic data.

Table 3 shows some words (nouns) taken from actual sample sentences in the corpus. The information in brackets { } shows the semantic group. These semantic groups can be used to create verb valence patterns. For example, we can create the following valence pattern for 'eat':

[{human, animal} '*ga*' {food} '*wo*' '*taberu*(eat)']

post-position	word			semantic class
が ('*ga*')	ラット (rat)	ウサギ (rabbit)	リス (squirrel)	{animal}
	子牛 (calf)	ベンガルヤマネコ (Bengal cat)	牛 (cow)	{animal}
	乳幼児 (infant)	妹 (younger sister)	貴族 (the nobility)	{human}
	私 (I)	子供たち (children)	お客 (customer)	{human}
を ('*wo*')	食べ物 (food)	日本米 (Japanese rice),	クルミ (milk)	{food}
	ドッグフード (dog food)	エサ（２） (bait; animal feed)	草 (grass)	{animal food}

Table 3: Semantic groups of Noun

6.2. Classifying Semantic Attributes

If you consider the semantic restrictions on the nouns that may be used with the word 折る '*oru* (fold)', you find that this has the rather broad description, [<concrete object> *'wo'* '*oru* (fold)']. However, by doing this, things that we cannot fold, such as '山 (mountain)' or '水 (water)' are included in the group. On the other hand, for a verb such as 'see' in [<concrete object> '*wo*' '*miru* (see)'], the semantic marker <concrete object> is ok. <concrete object> is ok as long as the object is tangible and we are able to visually see and recognize it. In other words, the level of semantic restrictions on the noun differs according to the verb. Once semantic groups are created, they can be arranged in a hierarchy. By arranging the semantic groups hierarchically, it allows us to select the appropriate level restricted by the verb. We feel the semantic group for the noun that would occur in the case of '折る (fold)' should be the semantic marker level valid for the verb rather than on a concept classification based on a traditional knowledge classification scheme.

Semantic Classification of Knowledge
concrete object
→ natural object
→ product
→ food
→ clothing
→ medicine
→ stationery
…

Within <stationery> and <food> (both sub concepts of concrete object) there are some things that can be folded and some things that cannot be folded. Instead of categorizing the object for 「折れる」'*oreru* (can be folded)' according to functionality or usefulness of the concrete object, we feel the object should be categorized according to a particular attribute of the concrete object's shape or form. Thus, in order to create semantic groups for nouns restricted by the verb, we are testing a classification structure based on attributes.

Semantic Marker（Attribute）Classification

concrete object
→（shape）
→ (solid)
{机 (desk), 箱 (box), カバン (bag)…}
→ (flat)
{紙 (paper),布 (cloth), ノート(notebook), スカーフ (scarf), スカート (skirt)…}

7. Conclusion

In this paper, we have presented how valence data has been compiled and how this data can be used as one means of resolving ambiguity that occurs in natural language processing.

8. References

EDR , 1996. The Japanese Co-occurrence Dictionary. *EDR Electronic Dictionary Version 1.5 Technical guide* Chapter7 pp. 1-49.

Kentaro Torisawa., 2001. Unsupervised Method for canonicalization of Japanese Postpositions. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* , pp.211-218.