A Knowledge Based Approach to Identification of Serial Verb Construction in Chinese-to-Korean Machine Translation System

Dong-il Kim*, Zheng-Cui, Jinji-Li**, Jong-Hyeok Lee
Department Computer Science and Engineering, Electrical and Computer Engineering Division,
Pohang University of Science and Technology (POSTECH)
and Advanced Information Technology Research Center (AlTrc)
San 31 Hyoja Dong, Pohang, 790-784, Korea
E-mail: {dongil, cuizheng, ljj,jhlee}@postech.ac.kr

Abstraction

In Chinese language processing, the recognition and analysis for serial verb constructions (SVCs) is a fascinating research topic. Chinese language researchers each may have a different definition and interpretation of SVC since the structure of SVC makes Chinese unique to other languages and contains complex semantic and pragmatic information. This paper proposes a formal definition of SVC and a knowledge based approach for the recognition of SVCs, which is adopted in TOTAL-CK, a transfer-based MT system from Chinese to Korean. The recognition process is carried out in two stages: the analysis stage classifies SVCs into general categories, and the transfer stage performs further classification for Korean transfer. Some evaluation result for each stage was also given with statistics of each category of SVCs

Introduction

Many Chinese language researchers have paid special attention to the so-called 'serial verb constructions (SVCs)', where two or more semantically or pragmatically related verb phrases or clauses are juxtaposed together without any functional marker. Because of different definitions and interpretations of SVCs among researchers, their categorizations differ according to researchers' viewpoints.

In a Chinese to Korean machine translation, the hidden relation of the serial verbs should be expressed with some function words from the target language viewpoint. Moreover, the conceptual scope of these function words is different from the scope of SVC categorizations that are classified based on the viewpoint of the Chinese language itself.

In this paper, we proposes a different categorization of SVCs defined by the contrastive analysis of the two languages, and also an SVC identification method that is adopted in a Chinese-to-Korean MT system, TOTAL-CK. The TOTAL (Translator Of Three Asian Languages: Chinese, Korean and Japanese) project has been conducted under a hybrid strategy with transfer-based and example-based engines.

1. Language Characteristics between Chinese and Korean

In this section, some contrastive analyses of the two languages are introduced for better understanding of an SVC sentence. Since Chinese is an isolating language, morphological or syntactic markers rarely appear in a sentence, while in an agglutinative language such as Korean, these functional markers are not an optional unit but an obligatory unit in a sentence.

An example is given in (1). Notice that the Korean alphabets are written with Yale Romanization in this paper.

---

* Also an assistant professor at Yanbian University of Science & Technology (YUST) Yanji, Jilin, China.

** Also a lecturer at YUST

(1) 他 开门 进去。(*ta kai-men jin-qu*)
Ku-nun mwun-ul yel-ko duleka-nta.
He-NOM door-ACC open-CON get in-PRENT-DEC.
He opens the door and enters (the room).

In the Korean sentence, *ko* is a connective particle, and also *nun*, *ul*, and *nta* denote a topic auxiliary particle, an object case particle and declarative terminative ending, respectively. All these functional markers should be decided in the Korean transfer stage. Specifically, we require a process to select one from the possible conjugational markers when a Chinese SVC sentence is transferred to its Korean counterpart.

2. Related Works

A SVC is studied among several researchers as different names. But the general syntactic form is (NP) V1 (NP) V2 (NP)[1]. The variance of definition for SVC comes from the different scope of interpretation for the sentence pattern.

We will introduce three typical researches to clearly outline our definition of SVC. The narrowest view of scope is suggested in (Lü, 1953). In his interpretation, V1 and V2 have the same subject and should be not coordinative, but it is difficult to decide which one is main or additional. Zhu (Zhu, 1981) includes all cases of Lü's and the possibility of adding an adjective to substitute for the second verb position. He also includes the case where an additional verb and a main verb are used, such as V+着 expression in V1 position, which indicates that V1 is additional and V2 is main. The broadest scope is proposed in (Li & Thompson, 1981). According to his interpretation, an SVC includes not only all the patterns noted above but also a pivot construction, a subject/object clause, and a coordinate clause, but excludes the pattern with an adjective in the V2 position. In this paper, the scope of SVC is almost same as Li's but the classification of SVCs differs slightly, detailing the categorization in chapter 4.

A few computational solutions to identifying SVCs have been proposed by some researchers. A formal description is shown in (Chan, 1998) using time lapse notation and the related definition. However, her method makes it difficult to computationally detect SVCs without the resources containing the deep level of analysis of each lexical, which is not obtainable in the current stage of language processing.

[1] V1 : first verb, V2 : second verb, NP : noun phrase.

3. Overview of TOTAL-CK System Architecture

As a typical transfer system, TOTAL-CK consists of three parts: Chinese analysis, dependency tree transfer, and Korean generation. The system architecture of TOTAL-CK is shown in figure 1. The design principles and the detail descriptions are given in (Kim *et al.,* 2002).

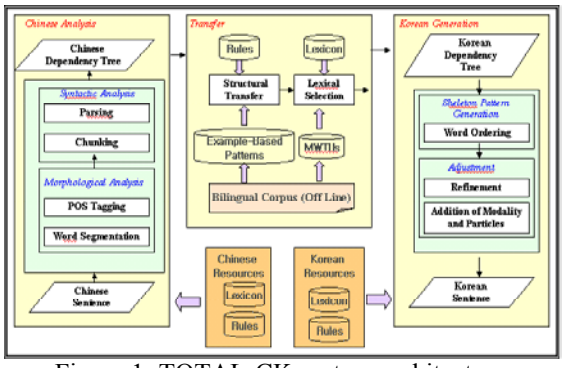

Figure 1: TOTAL-CK system architecture

4. Classification of Serial Verb Construction

In the previous chapter, we mentioned the syntactic format of SVCs which is NP V1 (NP) V2 (NP) and the different scope of definition of SVCs by the Chinese language researches. To outline the scope of SVCs, we define SVCs in terms of dependency relation such that V1 is the head of V2, or V2 is the head of V1. It is formally defined as follows:

Definition 1

Let N represent a set of nodes in a dependency tree, and W a set of words. Further Let V be a set of verbs, and P be a set of all parts of speech in Chinese. Then the functions: head, nw, and npos, are defined as below:

head(n) =hn where n∈N and hn is the head of n
nw(n) = w where n ∈N and w ∈ W
npos(n) = np where n ∈N and np ∈P

A definition of SVC is:

Given a node n such that npos(n) ∈ V and Head(n) = hn,
If and only if npos(hn) ∈ V and hn is the head of a given sentence then the sentence is a SVC.

The three sentences from the top of table 1 satisfy the given condition. Also the head of

the node is the sentence head, thus these must be SVCs. For the last sentence, nw(n) is 接入, and nw(Head(n)) is 总数 whose the POS is not verb and also whose the node is not the sentence head. Thus it is not an SVC.

Sentence	nw	nwh	SH	SVC
他开门进去。	进	开	Yes	Yes
我没想到你住在北京。	住	想	Yes	Yes
在这里停车犯法。	停	犯	Yes	Yes
各接入网络的总数已经超过1000个。	接入	总数	No	No

Table 1: Example of Testing SVC

Where nw: nw(n); nwh: nw(head(n)); SH :testing if head(n) is the sentence head ; n is a given node.

Our definition is employed to recognize a SVC in the Chinese analysis stage. First we describe the classification that is used in the Chinese analysis stage.

4.1 Categories in Chinese Analysis Stage

All dependency relations, which are detected by the above definition, are classified into five categories: separate events, object, subject, pivotal construction and descriptive clauses, based on the classification of Li (Li & Thompson, 1981).

4.1.1 Separate Events

The serial verb patterns classified by most researchers belong to this group where switching V1 to V2 provides us a different meaning. In addition, we add the case where transposing V1 to V2 provides us the same meaning in this group.

4.1.2 Object

If V2 is the main verb in an object clause or a object phrase then it belongs to this group.

4.1.3 Subject

If V1 is the main verb in the subject clause or subject phrase, it is assigned to this group.

4.1.4 Pivot

If the noun phrase between V1 and V2 is the object of V1 and the subject of V2, then it is a pivot construction.

4.1.5 Descriptive

If V2 describes the noun phrase between V1 and V2, then it is a descriptive SVC.

All categories of SVCs are shown in Table 2

The corresponding Chinese dependency relations to object, subject and pivot constructions also appear in the some research in Chinese language processing (Zhou & Huang, 1994) but the other two are not shown due to their different viewpoints.

The descriptive construction is directly able to be one-to-one mapped to the Korean counterpart. However the separate event SVCs should be further classified for Korean transfer since the separate event SVC is possibly mapped into sentences with several different Korean conjunctional particles. Thus, it is touched in the transfer stage.

Category	Example
Separate Event	我买票进去; 他在饭店吃饭喝茶。
Object	我没想到你住在北京。
Subject	在这里停车犯法。
Pivot	我们让他去北京。
Descriptive	我有一个姐姐喜欢看书。

Table 2: Examples of SVC category

4.2 Subcategories in Transfer Stage

The separate event SVC for each sense of Korean conjunctional particles is classified into the following subcategories: restrictive, quasi-coordinative, simultaneous, transitional, and circumstantial by the Korean language viewpoint.

4.2.1 Restrictive

The action of V2 is performed under the restriction given by V1. There are different types of restriction, such as space, group-related, causal, and instrumental. The examples are presented in table 3.

Sentence	V1	V2	R type
我的妹妹今天离开北京前往汉城[2]。	离开	前往	space
他代表山西省出席了座谈会。	代表	出席	group
我投票赞成第一个人。	投票	赞成	causal
这个图书馆利用计算机进行图书借阅管理。	利用	进行	tool

Table 3: Examples of Restrictive Separate Events

[2] The sentence can also be interpreted as purposive separate events. But it is included into a restrictive separate event SVC because it is impossible to detect the differences between restrictive and purposive, as this requires pragmatic level information

4.2.2 Quasi-Coordinative

In quasi-coordinative, two different cases exist. First, transposing V1 to V2 never causes a meaning shift of the sentence, named alternative. The other is that V1 and V2 are only sequentially related, called consecutive.

4.2.3. Simultaneous

In a simultaneous case, V1 and V2 occur at the same time.

4.2.4 Transitional

If the action of V1 is interrupted by the action V2, then it is transitional.

4.2.5 Circumstantial

When V2 occurs on the condition of the action of V1, then it is classified as a circumstantial case.

The examples for rests of the separate event are given in Table 4.

Type	Example
Q-Coordinative	他在饭店吃饭喝茶。(alternative) 他买票进去。(consecutive)
Simultaneous	他站着唱歌。
Transitional	我的弟弟开车出事了。
Circumstantial	不买别进。

Table 4: Examples of Separate Events

In restrictive, quasi-coordinate, simultaneous, transitional, and circumstantial separate event SVC Chinese sentences, all the above verbs are mapped into the corresponding Korean verb followed by the Korean conjunctional particle 'se', 'ko', 'un-chay-lo', 'taka' and 'myen', respectively.

5. Identification of SVCs

To recognize SVCs, we divide the identifying process into two stages. The general categories of SVCs are able to be found at the analysis stage and the subcategories of a separated event SVC are detected in the transfer stage.

5.1 Analysis Stage

To recognize the five general categories of SVCs, two resources are used: one is the Grammatical Knowledge Base of Contemporary Chinese (GKBCC) and the other is a verb list with valency information (VLVI) (Zhu *et al.*, 1995). Checking a verb in GKBCC allows us to simply detect a pivot SVC. The remainders of the other types of SVCs should be carefully handled. There are two possible ambiguous structures of SVCs

Case 1 : NP V1 V2 (NP2)
Case 2 : NP V1 NP1 V2 (NP2)

Where NP, NP1 and NP2 are noun phrases.

The algorithm for each case is illustrated in figure 2 and figure 3. In Figure 3, the test 'V1 takes NP & VP' means that the verb 偷听 can have a noun phrase or an object clause as an object. The test, 'satisfy valency' denotes that the second verb 喜欢 takes a human subject, and 外国人 can be the subject of the verb 喜欢, thus it is classified as an object case. For the other sentence, since 公园 cannot be the subject of the verb 锻炼, it is determined as a subject case.

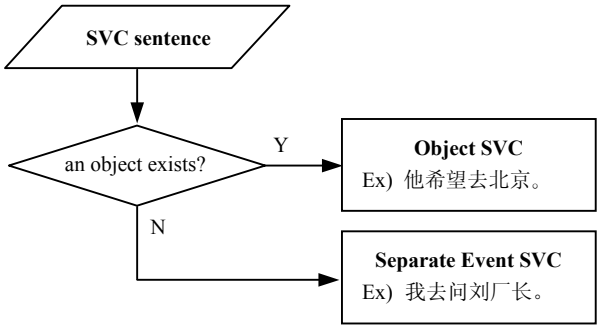

Figure 2: Algorithm of Detecting SVC for Case 1

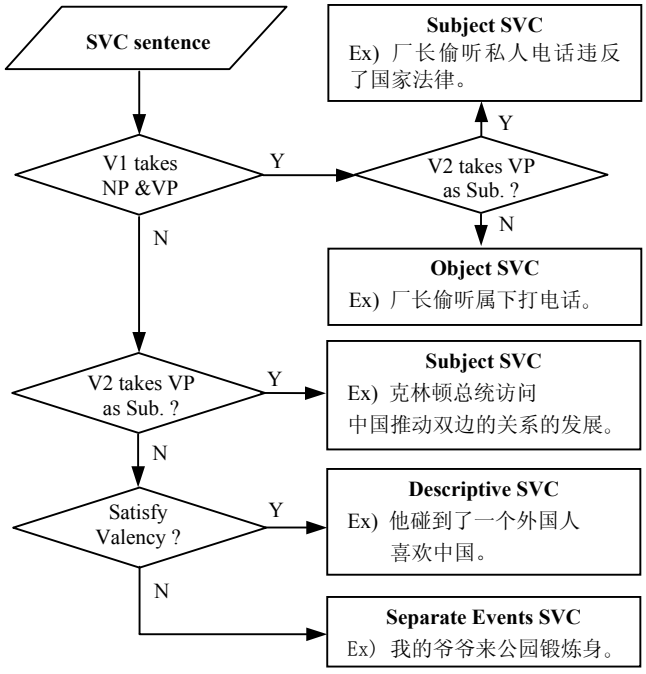

Figure 3: Algorithm of Detecting SVC for Case 2

5.2 Transfer Stage

The simultaneous separate events is easily recognized by the lexical (着) attached to the first verb. Also, we use a simple heuristic to detect the circumstantial separate events with the lexical pattern information.

The resource used in this stage is a Chinese thesaurus called Tongyi-ci-cilin (Mei, 1983). With the thesaurus the remainders of separate event SVCs are processed with great care. If V2 is related to the interrupt concept then the transitional separate events are assigned. The most difficult and frequently occurring cases are the restrictive separate events and quasi-coordinative separate event.

The key idea of using the thesaurus is based on the observation that the verb V2, if restricted by V1 makes it possible that the concept of V2 will also be restricted by the concept of V1. To complete the solution, we first define the relations: RSTV, RSTL and RSTM as follows:

Definition 2

We define the relations: RSTV, RSTL, and RSTM, as follows:

RSTV= {(V1,V2) where V1 and V2 are the first verb and second verb in a given SVC sentence and V2 is semantically restricted by V1 : (V1,V2) ≠ (V2,V1)}

RSTL= {(CL1,CL2) where CL1 and CL2 are the low level concept of the first verb and the low level concept [3] of second verb in the Chinese thesaurus, respectively, and CL2 is semantically restricted by CL1 : (CL1,CL2) ≠ (CL2,CL1)}

RSTM= {(CM1, CM2) where CM1 and CM2 are the middle level concept of the first verb and the middle level concept of second verb in the Chinese thesaurus, respectively, and ML2 is semantically restricted by ML1 : (ML1,ML2) ≠ (ML2, ML1)}

The relations RSTV, RSTL, and RSTM are not symmetric and not reflexive. Based on the definition we derive the following heuristics:

if (V1,V2) ∈ RSRV then (CL1,CL2) ∈ RSTL But if (V1,V2) ∈ RSTV then not always (CM1,CM2) ∈ RSTM.

All three examples from the top of table 5 satisfy the condition that, if (V1, V2) ∈ RSTV then (CL1,CL2) ∈ RSTL and (CM1,CM2) ∈ RSTM. If the condition is always true, then we use the middle-level concept relation for detecting a restrictive separate event in order to increase the applicability of our rules. Also, the data structure of RSTM is easily represented with an adjacent matrix with the size of 21*21 [4] (Sahni, 1998) where the matrix M is a square matrix, whose column and row are the middle-level concept, and if M(i,j) = 1 then concept j is semantically restricted by concept i, otherwise (i,j) ∉ RSTM.

Example	RSTV		RSTL		RSTM	
	V1	V2	CL1	CL2	CM1	CM2
他参加政府会议公开批评了我的谈话。	参加	批评	Hj20	Hi21	Hj	Hi
国家主席江泽民出席讲话。	出席	讲话	Hj20	Hj12	Hj	Hi
他给华大使带来一个大花篮表示祝贺。	带来	表示	Hj36	Hj14	Hj	Hi
他代表山西省出席了座谈会。	代表	出席	Hi17	Hj20	Hi	Hj

Table 5: Example of RSTV, RSTL and RSTM

However, the last example reveals that the condition is not always true since we have the result, both (Hi,Hj) and (Hj,Hi) ∈ RSTM. Thus, it violates the definition of RSTM. Hence, we may not directly use the middle-level concept adjacent matrix and the size of the low-level concept matrix is too large to be used.[5]

We come up with a solution of a frame with multi level concepts. The frame consists of three parts: the middle-level concept adjacent matrix, the low-level concept adjacent lists and the collocation serial verb list for detecting a serial verb that always appears together.

Our solution is that the exceptional cases are covered by either the collocation verb lists or the low-level concept adjacent list. The remaining frequently occurring cases are captured by the middle-level adjacent matrix. This leads to the sparse matrix of the low-level concept which

[3] The thesaurus consists of three levels of hierarchy. For example, H, Hj, and Hj20 correspond to the one of highest concept, the next narrow term called middle-level concept and the narrowest term called low-level concept, respectively.

[4] The number of verbs related middle-level concept in the Chinese thesaurus is 21.

[5] The number of verbs related low-level concepts in the Chinese thesaurus is about 500.

causes the adaptation of adjacent lists rather than an adjacent matrix for the low-level concepts.

The order of searching the frame is the collocation list, the low-level concept list and the middle-level concept matrix. In the collocation list, if V1 and V2 belongs to the collocation list of the restrictive separate events, such as 捉拿归案 or the one of quasi-coordinative, such as 立案侦察 then the sentence is assigned to a restrictive case or a quasi-coordinative case, respectively. In the low-level concept lists and the middle-level concept matrix, if matching succeeds, which means that V2 is semantically restricted by V1, then a restrictive case is assigned; otherwise, a quasi-coordinate case is detected[6]. The detailed process for identifying the subcategories of separate events is shown in figure 4.

6. Evaluation

We randomly selected 1000 SVC sentences from 1998 people's daily newspapers. The number of verbs in the sentence is two since our dependency parser is still being improved to detect the sentences with multiple embedding clauses. In table 6, the distribution of each type of SVC and the precision are shown.

Type	Frequency	Percentage
Separate events	402	40.2
Object	479	47.9
Subject	31	3.1
Pivot	39	3.9
Descriptive	1	0.1
Error	56	5.6 (Presicion:94.4%)
Total	1000	100

Table 6: Distribution of Categories of SVC

The precision is 94.4% and some of the errors occur from the tagger, thus some sentences are not SVCs. The rest of the errors result from missing information in the knowledge bases:

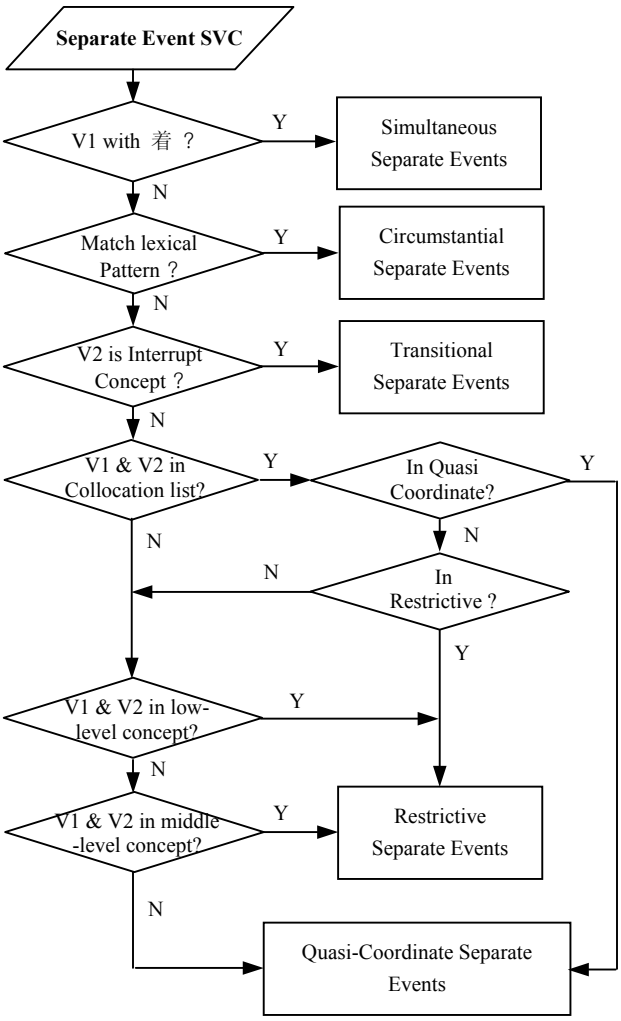

Figure 4: Detection Algorithm of Subcategory of Separate Event SVC.

GKBCC and VLVI. We need the complete list of verbs, which has a clause as a subject. These verbs in the list will be gradually collected in future works.

The evaluation table for the separate event SVCs is provided in Table 7.

Type	Frequency	Percentage
Restrictive	153	38.5
Quasi-coordinative	184	45.7
Simultaneous	33	8.2
Transitional	3	0.7
Circumstantial	12	2.9
Error	19	4.7 (Precision:95.3%)
Total	402	100

Table 7: Result of Separate Event

6 For a sentence 国家主席江泽民出席讲话 where the relation (Hj20,Hj12) is not in the low-level adjacent list, but (Hj,Hi) is 1 in the middle-level matrix, it is assigned to the restrictive case, while for the sentence 他代表山西省出席了座谈会 where (Hi17,Hj20) is in the low-level adjacent list, thus searching is stopped, it is assigned as a restrictive case. A sententence 他在饭店吃饭喝茶 do not satisfy all conditions, thus it is detected as Quasi-Coordinate.

The precision of identifying the category of separate event is 95.3%. The errors resulted from a circumstantial case since our heuristics is too restrictive to detect all cases, thus, it might be revised further, and since the low-level concept lists are not completed. The low-level concept lists will be continuously updated for increasing coverage in the tuning stage of the machine translation system.

Table 8 shows the distribution of the subcategory of restrictive separate events for Korean transfer.

Type	Frequency	Percentage
Space	86	56.2
Group-related	38	24.8
Causal	17	11.1
Instrumental	12	7.9
Total	153	100

Table 8: Category of Restricted Separated Event

In table 9, the frequency for each type of accessed resource is listed. Notice that most restrictive separate event SVCs are recognized in the middle-level matrix. The two cases in collocation are all the case of quasi-coordinative, thus, the total number is greater than 153.

Type of accessed resource	Frequency	Percentage
Middle-level matrix	121	78.0
Low-level list	32	20.6
Collocation list	2	1.29
Total	155	100

Table 9: Access Frequencies for Resource Type

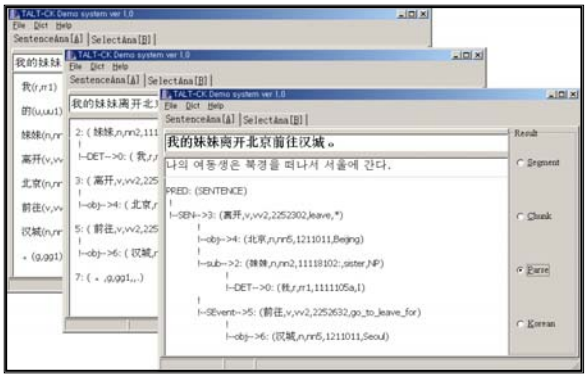

Figure 5: Demo system of TOTAL-CK

In figure 5, a demo system of TOTAL-CK is illustrated. For a given Chinese SVC sentence displayed in the top position of the right-most window, the corresponding Korean sentence is followed in the next row. The tagged results, the segment of chunking, and the Chinese dependency tree with indentation are shown in each window from left to right.

Conclusion and Future work

In this paper, we formally define serial verb constructions, and classified the SVC into several categories. These categories are related to the analysis stage and the transfer stage of TALK-CK. We provided a resolution algorithm detecting SVCs in each step. Finally, at each stage, a promising experimental result is shown.

Further research must help to better resolve the conditional separate event SVC and purposive separate event SVC.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center(AITrc).

References

Chan Y. W. (1998) *Formal Criteria for interpreting Chinese Serial Verb Constructions*. Communications of COLIPS 8(1) ,pp.13-29.

Kim D.I., Cui Z, Li J.J. and Lee J.H. (2002). *Resolving Structural Transfer Ambiguity in Chinese-to Korean Machine Translation*. 2002 International Conference on Chinese Language Computing, Taichung, Taiwan.

Li C N. , Thompson S A. (1981), *Mandarin Chinese: A functional reference grammar.* University of California Press, USA.

Lü S. X.(1953) *Yufa Xuexi (The Study of Chinese Grmmar)*. Bejing, Zhungguo Qingnian Press.

Mei J.J.(1983) *Chinese Thesaurus (Tong-Yi-Ci-Ci-lin)*. Shanghai Cishu Press. 1983.

Sahni, S.(1998) *Data structures, algorithms, and applications in C++*. Boston McGraw-Hill, USA.

Zhou M and Huang C. (1994) *Approach to the Chinese Dependency Formalism for the Tagging of Corpus*. Journal of Chinese Information Processing, 8/3, pp. 35-52, 1994

Zhu D.X.(1981), *Yufa Jiangyi (Lectures on Chinese Grammar)*,Bejing, Xiangwu Press.

Zhu X. F, Yu S. W and Wang H. (1995) *The Development of Contemporary Chinese Grammatical Knowledge Base and its Applications.* Communications of COLIPS, 5/1-2