Building Carefully Tagged Bilingual Corpora to Cope with Linguistic Idiosyncrasy

Yoshihiko Nitta*, Masashi Saraki* and Satoru Ikehara**
*Nihon University College of Economics, **Tottori University College of Engineering
*1-3-2 Misaki-cho, Chiyoda, Tokyo 101-8360 Japan, **Tottori, 680-8553 Japan
E-mail: nitta@eco.nihon-u.ac.jp, saraki@st.rim.or.jp, ikehara@ike.tottori-u.ac.jp

Abstract

We illustrate the effectiveness of medium-sized carefully tagged bilingual core corpus, that is, "semantic typology patterns" in our term together with some examples to give concrete evidence of its usefulness. The most important characteristic of these semantic typology patterns is the bridging mechanism between two languages which is based on sequences syntactic codes and semantic codes. This characteristic gives both wide coverage and flexible applicability of core bilingual core corpus though its volume size is not so large. A further work is to be done for grasping some intuitive feeling of pertinent coarseness and fineness of patterns. Here coarseness feeling is concerning the generalization in phrase-level and clause-level semantic patterns and fineness is concerning word-level semantic patterns. Based on this feeling we will complete the core tagged bilingual corpora while enhancing the necessary support functions and utilities.

1. Introduction

We propose to give an overview of the results of our work during these five years by CREST work project which is jointly processed by Tottori University, Nihon University, Niigata University and Gifu University of Japan. This project at first stage collected carefully the well-written bi-lingual sentences and texts mainly from ordinary life-style writings such as newspapers, magazines, books, commercial pamphlets and various popular articles. Here "bi-lingual" means "between Japanese and English".

By comparing carefully a pair of sentences written in different languages, we can grasp lots of linguistic idiosyncratic phenomenon. The project collected 2 million bilingual sentences from various ordinary life fields, then, which are carefully filtered to obtain well written representative 215 thousand sentences. Using these bilingual sentences as a kind of core sentences, we put part-of-speech-tags, coarse semantic markers associated with case-frame markers.

Semantic markers (200 entries) are constructed as a super set of semantic categories (3000 entries) defined by Japanese Vocabulary Survey (1997 Ikehara et al.). Starting from these tagged core bilingual sentences, we made linguistic refinement (including corrections, modifications and/or rephrasing) together with generalization and specialization by introducing some semantic- and syntactic- codes and variables.

Finally we obtained precisely tagged core bilingual sentences. Let us refer to these results as "semantic typology patterns".

The objectives of project, however, is not to construct a large semantically tagged bilingual core corpus but rather to suggest or even prove that the application and/or implementation of carefully tagged medium-sized core bilingual corpus can play an important role for resolving various idiosyncratic matters occurred in various NLP applications such as in MT, TG, IR, IE and QA.

2. Motivation and Objectives

The motivation of this research is to facilitate the automatic translation between two different (somewhat far distant) languages, cross-lingual information retrieval, text generation and question answering, which are currently hindered by much of idiosyncratic matters.

Nowadays prevailing techniques for attacking linguistic gap or idiosyncrasy are statistical methods based on huge volume of corpora, but still which cannot cope with data sparseness problem. Because the semantic and syntactic difference (or variance) between two different languages is too large to resolve only by brute force statistical methods.

Our hypothesis and/or assertion are that before going into the implementation of massive corpus-based statistical methods, we should construct carefully some medium-sized well-designed bilingually tagged corpora as a kind of core (seed) linguistic knowledge.

3. Background

3.1. The CREST MT Research 01-05

This research has been constructing carefully designed annotated bi-lingual (English-Japanese) corpus as its byproduct. The main purpose of the CREST research is to develop the theory, methodology and experimental implementation machine translation system, which is strongly naive and intuitive semantics-oriented.

The research results will facilitate both future machine translation methodology and language educating/learning methods. The research products will be open (free) to researchers and ordinary language users.

The purpose of the CREST research is to investigate the method to overcome the performance limit of classical transfer-base machine translation systems, which essentially based on the syntactic structure transformation. The corpus-base or example-base machine translation, which is the nowadays-mainstream translation paradigm,

cannot clearly present the linguistic meaning because of its inherent black box processing property. This kind of opacity inevitably leads to the naive feeling that the machine is not carrying out for the human-like translation tasks.

The classical transfer-base machine translation has sound and transparent mechanism. Our result is that if we adopt the carefully designed linguistic patterns for the objects of transfer, we can increase drastically the translation performance. The newly designed patterns can represent not only the surface sentential or syntactic patterns but also the semantic patterns such as predicate augment structure. The pattern transformation is carried out based on the concept of "semantically equivalent transformation", which is a mechanism for preserving logical semantics of each sentence. Some translation examples are shown together with the internal translation steps.

The NLP research topics relating the CREST research are machine translation, pattern description, regular expression, semantically equivalent conversion, syntactic structure, semantic analysis, structural transfer and linguistic idiosyncrasy.

3.2. Problems Addressed by CREST

The research on the technique that analyzes natural language expression using the knowledge of computer science is pursued from various application purposes, such as language understanding, question answering, text generation, information extraction, and automatic summarization. Moreover, apart from the application purpose, based on the interest of purely cognitive science study, human's language understanding capability and language generation capability are studied using the help of the knowledge and means of computer science (Marcus 1980; Winograd 1983).

The CREST project investigates the human's natural language understanding abilities from the viewpoint of computationally formalized linguistic models. Needless to say, the human's language understanding tasks are ranging widely; here we put our focus on the language translation tasks, especially English into Japanese translation. Moreover, as for the linguistic formalization, we put our attentions on the internal language models used in a classical (transfer-base) machine translation system (=MTS). The reason why we adopt the classical MTS is that their internal language models are intuitively and cognitively well-expressed humans language manipulating abilities and behaviors. In the case of corpus-base (or example-base) MT, or statistical MT, and various speech (or verbal) MT, the internal language models are far distant from the humans' intuitive models. They are purely mathematical and computational formulae, which are not suitable for cognitive language studies.

If we get fairly successful results, the inevitably we will reach the reevaluation of the classical MTS which will surely make a strong trigger to resuscitate classical transfer-base MTS.

An abstract specification of the mechanism of classical (English-into-Japanese) MTS may be summarized as follows (Kuno et al. 1962; Lehmann et al. 1980; Robinson et al. 1980; Lehrberger 1978; Nitta 1982; Nitta 1984).

(1) To segment the input English sentence into the sequence of phrase elements and/or clausal elements,
(2) To assign grammatical marks (such as part-of-speech and syntactic roles) to each segmented elements,
(3) To embed the dependency relation and/or mother-daughter relations so as to construct the sentence patterns from the marked segment sequence above,
(4) To transform the structured pattern above into the new pattern that reflects the word order (and composition) of target language,
(5) To replace the each elements by the adequate target language components,
and
(6) To make a fine adjustment of the surface sentence pattern(s) above so as to form natural and eloquent target language sentence(s).

Traditionally the steps (1) ～ (3) are often called "analysis phase", the steps (4) is called "transfer phase" and steps (5) ～ (6) are called "generation phase".

In short, the essence of the classical machine translation is that the skilful transformation of sentential patterns which are the structured sequences of various linguistic or grammatical symbol sequences so as to obtain the word or phrase sequences of the target language while preserving the original meaning (=semantics) of input (=source language) sentences. These generalized and abstracted translation steps are clearly reflect well the human's (i.e. somewhat lousy foreign language beginners') translation process.

Consequently the problems addressed here are to design carefully the adequate linguistic patterns so as to grasp the semantics of both source- and target-languages, and to construct the skilful transformation mechanism to transfer the source patterns into target patterns so as to obtain adequate (high quality) translations. To sum up, our research goal is to establish the "semantically equivalent pattern transformation" (Nitta 2002a; Nitta 2002c).

3.3. The Intermediate Results of the CREST MT Research

We have got some concrete evidence that if we adopt the carefully designed linguistic pattern description method(s), we can increase substantially the translation quality of classical transfer-base machine translation while preserving the intuitive and transparent properties of them.

The pattern description of source- and target-sentences are defined and controlled by the "meta-patterns". Meta-patterns define various linguistic components such as word elements, phrasal elements, clausal elements, syntactic roles, pending modifiers; attribute sequence of noun and noun phrase, governing type of verb and verb phrase, and so on. The internal structural pattern transformation process and language generation phases are also defined and controlled by meta-pattern descriptions.

More extensive adaptation of word and phrase semantics and/or semantic descriptions is the further subject of our research. Currently we are trying to adopt

the some logical (predicate-type) description of clausal semantics based on the word semantic lexicon system (Ikehara et al. 1997). The key point of logical semantics specially featured for MTS is the predicate logics to grasp logical relationships among phrases or clause. In terms of these logical semantics, classical (pattern-base) MTS can identify efficiently adequate pattern(s) correspondence for language transfer.

The relevancy study of our proposed translation method and pattern description method are the next sub goal of our research.

4. Pattern Transformation

The language learning assistance is performed as the flow of linguistic pattern transformation. Language learners can learn and recognize some tacit language manipulation recipe during tracing the machine providing pattern transformation process. In the following, we will show the semantically equivalent pattern transformation:

(1) Word-level Pattern Transformation
Input (Japanese) sentence: *Ukkari shite teikiken wo ie ni wasure tekita.*
↓ (Analysis)
Japanese pattern: #1(N1 *ha*) /V2 *te* /N3 *wo* /N4 *ni* /*tekita*(V5).
↓ (Transfer)
English pattern: I t was so AJ(V2) as to V5 *poss*(N1)N3 at N4.
↓ (Generation)
Output (English) sentence: It was so careless as to leave my season ticket at home.

(2) Phrase-level Pattern Transformation
Input (Japanese) sentence: *Goukaku no shirase wo kiite kanojo no kao ha akaruku natta.*
↓ (Analysis)
Japanese pattern: VP1 *te* /#1 (N2 *no*) /N3 *ha, past*(VP4).
↓ (Transfer)
English pattern: When N2 *past*(VP1), #1(*poss*(N2))N3 *past*(VP4).
↓ (Generation)
Output (English) sentence: When she heard she had passed the examination, her face brightened up.

(3) Clause-level Pattern Transformation
Input (Japanese) sentence: *Kore ha kiwamete yudoku dearu-node, shiyou ni atatteha junibun ni chui shinakuteha naranai.*
↓ (Analysis)
Japanese pattern: CL1 *node*, N1 *ni atatteha* /*must*(VP2)
↓ (Transfer)
English pattern: *so+that*(CL1, *passive(must*(VP2))) with *poss(subj*(CL1) N1)
↓ (Generation)
Output (English) sentence: It is significantly toxic so that great caution must be taken with its use.

Note that in the above transformations, a kind of bi-directionality holds; that is to say, if we reverse the transfer direction, we get English-into-Japanese translation. The vital issue for machine translation is to achieve the transfer process as accurate as possible. For this purpose, we devise a mechanism to refer semantic typology pattern(s) [Ikehara et al. 02].A brief sketch of the semantic typology patterns is as follows:

'wasureru'('wo': physical-object, 'ni': place) = leave(physical-object in /at place)
'wasureru'('wo' :information/knowledge) = forget(information /knowledge)
'akaruku-naru'('ga': place) = lighten /brighten(place)
'akaruku-naru'('ga': mood /mental-state) = become cheerful(mood/mental-state)
'akarui'('ni': information /knowledge) = be familiar with/be acquainted with(information /knowledge)
〔clause incl: 'adjectival / adverbial'〕+ 'dearu-node' + 〔clause2〕= clause-adjectival / adverbial ; so adjectival / adverbial that +〔clause2〕

Roughly speaking, the key idea of semantic typology pattern is that: the concise and effective knowledge that gives preferable combination of "noun with its semantic category" and "predicate phrase (such as verb phrase) class" and/or that of "clause with its attribute" and "conjunction class" can increase the quality of 'transfer' drastically. So far we have finished the description of semantic patterns: 130,000 entries for word-level, 100,000 entries for phrase-level and 20,000 entries for clause-level.

5. Semantic Typology Pattern

The Semantic Typology Knowledge Base provides contextual condition that controls the intra- or inter-sentential structure, thus can gives an effective means for appropriate foreign language manipulation recipes.

5.1. Semantic Typology in Word Level

* Evaluation and Judgment:
～[の]は AJ
～が AJ
～することは/が AJ
彼が怒るのは至極当然だ。 → It is quite natural that he should get angry.
日本は有色人種の弁護の任に当たるのが 当然である。 → Japan ought to stand up for the colored races.
その２つの三角形の面積は等しい。(その２つの三角形の面積は同じだ。)
→ The areas of those two triangles are the same.
(Those two triangles have the same areas.)
豚は有蹄類と考えることが正しい。→ The pig is properly regarded as an ungulate.

5.2. Semantic Typology in Phrase Level

* Evaluation and Judgment:
Phrasal level analysis is superficially almost same as the case of word level, except for the part-of-speech conversion between noun and verb. Here we will give only one example. More accurate analysis is postponed to

the next paper.
子どもには過大な期待をかけないほうがいい。→ It is best not to expect too much of your children.

5.3. Semantic Typology in Clause Level

In the case of clausal level, almost all the semantic typologies have some connections with semantic relations between two consecutive clauses.
*Negative Condition and Negative Result:
先週苦情処理を終えたばかりなのに、今週また別の苦情が寄せられました。
→ I just had finalize a claim last week, but I have got another one this week.
Japanese connective expressions of the similar kind are as follows:
あげくに、　あげくの果て、にもかかわらず、というのに、それどころか、とたんに、くせに、けれども、からには、ところが、どころか、であるのに、とはいうものの、

6. Some Findings from Tagged Bilingual Corpus

In Japanese many compound sentences are of form "～ shite ～ suru" which means "VP + VP", where VP stands for verb phrase. VP may be replaced by V (=Verb). Thus if we can classify Japanese compound sentences of shite-form effectively, which will facilitate the learners of English compositions and/or Japanese-into-English translation, even foreigners who intend to learn Japanese basic sentential patterns.

Japanese shite-form sentences are classified into following three syntactic categories (Saraki and Nitta, 2005):

(1)
(1.1) V1 して V2 verb conjunction only
(1.2) V1 ようにして V2 comparing situation
(1.3) V1 ようとして V2 intention
(1.4) V1 させて V2　causative verb
(1.5) V1 られて V2　passive verb
(1.6) V1 ないで V2 negation

(2)
(2.1) V1 して V2 する
(2.2) V1 して VP2 する
(2.3) VP1 して V2 する
(2.4) VP1 して VP2 する

(3)
(3.1) V (P) 1 して V (P) 2
(3.2) V (P) 1 して V (P) 2 して V (P) 3

Japanese shite-form sentences are classified into following four semantic categories (Saraki and Nitta 2005):

A. Collateral condition

(1) Agent condition

(1.1) Posture change
太郎は小首をかしげてしきりに考えていた。
Taro was thinking hard with his head on one side.
(1.2) Put on-off
父はグレーの背広を着て出かけた。
Father went out in his gray suit.」
(1.3) Carrying
ずっしり重いかばんを下げて兄が出張から帰ってきた。
My elder brother came home from his business trip carrying a heavily packed bag.

(2) Mental condition

(2.1) Inner mental action
彼はあわてて彼女を押しのけて行った。
He hurriedly pushed past her.
(2.2) Exposed mental condition
花子は口許に微笑を浮かべてわたしたちを迎えてくれた。
Hanako welcomed us with a smile on her lips.
(3) Agent's action
その犬は鼻をくんくんさせて食べ物を探し回った。
The dog sniffed around for food.
(4) Collateral condition
(4.1) Agent's condition
彼は先に立って歩いた。
He walked in front.
(4.2) Condition when main event (main predicate) occurs
義雄はよくテレビをつけっぱなしにしてうたた寝している。
Yosio often dozes off with the TV on.
(4.3) Conditions expressed by similar events or metaphor
太郎と花子は人目に立たないようにして会っていたものだ。
Taro and Hanako used to meet secretly.

B. Temporal condition

(1) Temporal circumstances
しばらくして手に痛みを覚えた。
After a while, I felt a pain in my hand.
(2) Temporal succession
母はスープの味見をして塩を加えた。
My mother tasted the soup, and then added salt.
(3) Temporal processing flow
その実業家は十分に足固めをして新しい事業に取りかかった。
That businessman made all necessary preparations before embarking on a new enterprise.
(4) Moment of action
多くの難民が脱走しようとして撃たれた。
Many refugees have been shot while making a bolt for freedom.

C. Originating condition

(1) Cause（reason of involuntary event occurrences）
その箱を持ち上げようとして梅子は肩の筋を違えた。
Umeko strained her shoulder lifting the box.

(2) Reason

(2.1) Reason of agent's voluntary action
太郎は欲が出て、失敗した。
Taro failed because he was too eager.
(2.2) Agent's judgment

危険が起こるかもしれないと考えて彼はあとに残った。
He stayed behind in view of possible danger.
(2.3) Bases of judgment (of inexplicit agent)
安全性、経済性の両面から考えて、このストーブを買った。
I have bought this stove after considering both economy and safety.
(2.4) Reason why agent is affected by other voluntary actions
彼は信号無視をして重い罰を受けた。
He was severely punished for running a red light.

(3) Objective-oriented cause

(3.1) Subjective intention
父はこちらを振り向かせようとして空咳をした。
My farther gave a dry cough to make them turn and face this way.
(3.2) Presentation of objective
多くの候補者が市長の椅子をめざして選挙運動中です。
Many candidates are campaigning for the mayor ship.

(4) Methodological cause

その国会議員は父を買収して何も言わせないようにした。
The congressman bribed my farther to say nothing.

(5) Condition

あなた無くしては一日も生きていられない。

D. Parallel

二人は芝生の上に仰むけになって寝ころんだ。
The two lay on the lawn face up.

7. Code System and Utility Functions

Some comments on Code System and Environmental Utility Functions are as follows.

Semantic typology patterns are composed of word-level patterns, phrase-level patterns and clause-level patterns, whose sizes are 123 thousands, 80 thousands and 12 thousands respectively. Here phrase-level and clause-level are a kind of semantic generalization of word-level semantics.

To facilitate various NLP applications that utilize our semantic typology patterns, we developed specially tailored morphological analyzers, sentence pattern analyzers and pattern matchers which select the most possible patterns corresponding to given sentences. Almost all these utility programs are coded as a kind of FSM (=finite state machine), that is, almost all the patterns are treated as somewhat enlarged regular expressions. These simple formulations contribute to the highly efficient computational processing.

Typical meta-symbols to describe patterns are variables (in total 15 kinds; word-level variables are of 9 kinds, phrase-level variables are of 5 kinds, clause-level variable is of 1 kind). We also developed program functions to treat part-of-speech tags, tense-aspect-mode tags, sequence of patterns, constituents ordering, wild-card symbols, empty symbol, etc.

8. Experiments

So far we have made large scale experiments to examine feasibility and effectiveness of our "semantic typology patterns". Here we would like to describe these experiments and outcomes. We began with annotating 215 thousand "patterns" with semantic category codes, syntactic codes (including part-of –speech tags), and some meta-symbols for facilitating regular language like treatments. Vital codes are case-markers and semantic category codes.

Our case marker code system is a little bit enhanced version in comparison with traditional Fillmore-style ones; it is composed of items that play a role in some logical and idealized macro-real-world. The semantic category code system is also a vital issue for successful language processing generally. Ours is a kind of super set taken from well-designed and widely used semantic classification categories (3000 entries) defined in Japanese Vocabulary Survey (1987) with some modifications. The modification is done mainly from a viewpoint of current common sense.

Feasibility tests (experiments) are done in two applications. One is concerning machine translation (MT) and the other is concerning text generation (TG).
As for MT, we developed a kind of transfer-base simple MT system using a regular language processing system (which we call extended pattern matchers) and utilizing our "semantic typology patterns" as a guiding linguistic knowledge. Evaluation experiments were done over 10 thousands sentences extracted randomly from our initial bi-lingual corpus covering more than 2 million sentence pairs. The result is that 75% of the outputs are of comparable quality as in original humans-hand translations. This shows the effectiveness of our semantic typology patterns, though some more extension efforts are necessary for really practical applications.

As for the other experiment on TG, we have devised a kind of generator for proverb-like and/or verse-like sentences. This system works when given some seed concepts as a list of keywords. Though an objective evaluation is a sort of tough problem, but still it has successfully revealed a substantial effect of our semantic typology patterns.

9. Conclusion

We conclude that medium-sized carefully designed linguistic pattern description, that is, "semantic typology patterns" in our term, is indeed feasible and effective to compose effective the knowledge-base for various natural language applications which require the concrete means to solve semantic idiosyncrasy problems.

In this paper we have only shown the typical examples obtained from our MT research activities. Further work is needed to enlarge and refine our semantic typology patterns together with more pertinent tagging and generalization works.

Most vital issues of this further work are to grasp some intuitive feeling of pertinent coarseness and fineness of patterns. Here coarseness feeling is concerning the generalization in phrase-level and clause-level semantic patterns and fineness is concerning word-level patterns. Also more precise investigation is necessary to build fine-grained coding systems such as in semantic categories and case markers together with supporting environment utility functions.

10. Acknowledgements

We would like to express our sincere thanks to the foundation of 21st Century CREST, Core Research for Evolution Science and Technology of Japan for providing financial support.

11. References

Bentivogli, L. and Pianta, E. (2005) Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: the MultiSemiCor Corpus. *Natural Language Engineering* 11(3):247-261

Hornby, A. S. (1978) *Guide to Patterns and Usage in English.* Oxford University Press (Japanese Translation: Kenzo Ito.1978. *An English Model, Usage*. Oxford University Publication Office)

Ikehara, S. et al. (1997) *Japanese Lexical Compendium*, Iwanami Shoten, Tokyo, Japan

Kinyon, A. (2001) A Language-Independent Shallow-Parser Compiler, *Proc. 39th ACL Ann. Meeting (European Chapter):*322-329

Lehrberger, J. J.(1978) *Automatic Translation and the Concept of Sublanguage*, Groupe de Recherché en Tradution Automatique (TAUM), Universite de Montreal, Canada

Lehmann, W. P. (1980) *The METAL System*, Linguistic Research Center, University of Texas, Texas, USA

Marcus, M. P. (1980) *A Theory of Syntactic Recognition for Natural Language*. The MIT Press

Mihalcea, R. and Simard, M. (2005) Parallel Texts. *Natural Language Engineering* 11(3):239-246

Moon, R. (1987) The Analysis of Meaning, in: (Sinclair (ed.), 1987) Chapter 4:86-103

Nakamura, Y. (1983) *How far can we go in translation*? Japan Times, Tokyo, Japan

Nitta, Y. et al. (1982) A Heuristic Approach to English-into-Japanese Machine Translation. in: J. Horecky (ed.).*Proc.COLING 82 (at Prague) (=Proceedings of the 9th International Conference on Computational Linguistics)*, North Holland Publishing Company: 283-288

Nitta, Y. et al. (1984) A Proper Treatment of Syntax and Semantics in Machine Translation, *Proc. of COLING 84 (at Stanford)) (=Proceedings of the 10th International Conference on Computational Linguistics),* Association for Computational Linguistics: 159-166

Nitta, Y. (1993) Referential Structure: A Mechanism for Giving Word-Definitions in Ordinary Lexicons. in: *Language, Information and Computation*, LSK (Linguistic Society of Korea)

Nitta, Y. (2002a) A Study of Semantic Typology Patterns and their Transformations, *Economic Review of Nihon University*, 71(4) Nihon University, Tokyo:131-155

Nitta, Y. (2002b) Problems of Machine Translation: From a Viewpoint of Logical Semantics, *Economic Review of Nihon University*. 72(2) Nihon University, Tokyo: 23-42

Nitta, Y. (2002c) A Study of Descriptive Language for Sentence Patterns, *Economic Review of Nihon University*. 72(3) Nihon University, Tokyo: 35-59

Saraki, M. and Nitta, Y. (2005) The Semantic Classification of Verb Conjunction in the "Shite" Form, *Proceedings of Spring IECEI Conference*, IECEI Japan

Slocum, J. (1985) Machine Translation---Its History, Current Status and Future Prospects, *Computational Linguistics*, 11(1)