Toward Automatic Compilation of Phrasal Thesaurus

Atsushi Fujita **Satoshi Sato**
Graduate School of Engineering, Nagoya University
{fujita,ssato}@nuee.nagoya-u.ac.jp

1 Introduction

Thesaurus, which links between linguistic expressions (or concepts) based on various semantic relations, is one of the most fundamental semantic resources in a broad range of NLP tasks. A lot of work has been carried out relying on thesauri, such as WordNet (Miller, 1995) and automatically created versions of it. The entries of most existing thesauri are either single words or word sequences including phrasal verbs and canned phrases. However, they may not be almighty in dealing with meaning, because meaning of polysemous word is determined only when it is used in some context, and meaning of phrase, clause, and sentence is determined based not only on those of its constituent words but also on its construction.

For a more precise semantic computing, we have proposed *phrasal thesaurus*, which regards phrases as entries (Fujita et al., 2007). While the term "phrase" generally refers to word sequences, such as phrasal verbs and canned phrases, in our study, the notion also includes predicate phrases those involve complements. Among various types of semantic relations between phrases, we have been addressing mainly paraphrases and textual entailment.

This paper describes the direction and current status of our study on compiling phrasal thesaurus.

2 Thesaurus of Predicate Phrases

Combination of content words and various constructions coerce us into handling an enormous number of expressions than word-based thesaurus. In addition, single concept can be conveyed by various constructions (Fujita, 2008). Our strategy to attain the coverage of paraphrases of predicate phrases is to divide them according to their productivity and required knowledge, and then separately develop resources to compute them.

Non-productive paraphrases

Synonym pairs (e.g., "comfort" $\Leftrightarrow$ "console") and idiom/literal paraphrases (e.g., "kick the bucket" $\Leftrightarrow$ "die") are typical examples of non-productive paraphrases. As they cannot be represented with abstract patterns, a huge amount of fully lexicalized paraphrase pairs should be compiled into a dictionary statically to realize this class of paraphrases. State-of-the-art corpus-based techniques for acquiring paraphrases are beneficial to its compilation. In fact, most of the paraphrases that previous work has collected are classified into this class.

In our study, we are compiling a idiom/literal paraphrase dictionary. So far, we have compiled a list of basic Japanese idioms based on five dictionaries for human, in order to set a goal regarding its scale; the resultant list consists of 3,629 entries (Sato, 2007). The next step is to collect the counterpart, i.e., literal phrase, for each basic idiom. While candidate literal phrases can be extracted for a certain portion of idioms from explanatory sentences in dictionaries for human, we will also apply automatic acquisition methods in order to complementarily attain the coverage.

Productive paraphrases

In contrast to non-productive paraphrases, it seems reasonable to represent the knowledge for productive paraphrases, such as voice/case alternation, nominalization and light-verb construction, with abstract patterns. For example, the phrasal pair (1a) can be represented with the pattern (1b).

(1) a. Employment shows a sharp decrease
$\Leftrightarrow$ Employment decreases sharply
b. X_{Noun} show a $Y_{Adjective}$ Z_{Noun}
$\Leftrightarrow$ X_{Noun} $verb(Z_{Noun})$ $adverb(Y_{Adjective})$

However, those patterns are not capable of preventing incorrect instantiations of phrasal paraphrases, because their applicability conditions, such as restrictions for variable slots (e.g., $Y_{Adjective}$ and Z_{Noun}), tend to be underspecified. This is fatal, particularly in case of generating paraphrases; for example, the following incorrect paraphrases are generated from the pattern (1b).

(2) a. Statistics show a gradual decline
$\not\Leftrightarrow$ Statistics decline gradually
b. The data show a specific distribution
$\not\Leftrightarrow$ *The data distribute specifically

Yet, this descriptive approach guarantees a certain degree of equivalence by exploring paraphrase instances based on transformation patterns and lexical functions, such as $verb(Z_{Noun})$ and $adverb(Y_{Adj})$.

On the basis of this recognition, we have been examining the following generate-and-test method, particularly targeting at Japanese:

Step 1. (Over-)generate syntactic variants based on syntactic transformation and lexical derivation.

Step 2. Measure how the pair of phrases is likely to be grammatical and correct as paraphrases by an empirical method.

In the first step, syntactic variants are generated using the following three sorts of linguistic knowledge (Fujita et al., 2007):

- Transformation patterns that give skeletons of syntactic variants, like (1b)
- Generation functions that generate a set of the simplest phrases from 0-2 content words
- Lexical functions whose back-end is ENJI: a Japanese Lexical Derivation Database

While the first two resources are fully handcrafted, the last one is semi-automatically compiled based on affix patterns, such as "S-i:Adjective $\Rightarrow$ S-mi:Noun" for "*amai* (be sweet)" and "*amami* (sweetness)." The database has been enlarged and cleaned up after (Fujita et al., 2007); consequently it consists of 4,814 pairs of cross-categorial lexical derivatives (3,525 trees containing 8,265 words).

Then, in the second step, each pair of automatically generated phrasal paraphrases are assessed against the following criteria that a correct pair of phrasal paraphrases must fulfill (Fujita and Sato, 2008a; Fujita and Sato, 2008b):

Criterion 1. Semantically equivalent

Criterion 2. Substitutable in some context

Criterion 3. Grammatical, respectively

To quantify how the given pair of phrases satisfies the above criteria, we have examined an empirical model which combines structured N-gram language models and distributional similarity measures.

Through a series of experiments, our approach has achieved promising results by coupling constituent similarity based on descriptive knowledge and contextual similarity computed empirically.

3 Future directions

While we described the motivation and current status of our study on compiling phrasal thesaurus particularly focusing on predicate phrases, we are also concerned with paraphrasing of functional expressions. Based on TSUTSUJI: a dictionary of Japanese functional expressions (Matsuyoshi and Sato, 2008), we are exploring the multi-word functional expressions, and the interaction between predicate phrases and functional expressions.

References

Atsushi Fujita, Shuhei Kato, Naoki Kato, and Satoshi Sato. 2007. A compositional approach toward dynamic phrasal thesaurus. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (*WTEP*), pages 151–158.

Atsushi Fujita and Satoshi Sato. 2008a. Computing paraphrasability of syntactic variants using Web snippets. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing* (*IJCNLP*), pages 537–544.

Atsushi Fujita and Satoshi Sato. 2008b. A probabilistic model for measuring grammaticality and similarity of automatically generated paraphrases of predicate phrases. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pages 225–232.

Atsushi Fujita. 2008. Survey on automatic paraphrasing. http://paraphrasing.org/.

Suguru Matsuyoshi and Satoshi Sato. 2008. Automatic paraphrasing of Japanese functional expressions using a hierarchically organized dictionary. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing* (*IJCNLP*), pages 691–696.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Satoshi Sato. 2007. Compilation of a comparative list of basic Japanese idioms from five sources. In *Information Processing Society of Japan SIG Notes, NL-178-1*, pages 1–6. (in Japanese).