The Frames-and-Constructions Approach to Paraphrase

Yoko Hasegawa, Russell Lee-Goldman, Kyoko Hirose Ohara
Michael Ellsworth, and Charles J. Fillmore

Paraphrasing is a pivotal concept in diverse academic fields, e.g., language education, lexicography, natural language processing, and translation studies. Nevertheless, theoretical investigation of this important field is surprisingly scarce. This paucity is to a great extent due to the elusive nature of the notion *paraphrase*. While it is possible to define the term strictly as "a relationship of mutual entailment that holds between two statements expressed by two sentences," in every-day use, *paraphrase* is more likely to refer to relationships broader than mutual-entailment. For example, under normal circumstances, (1a) can be considered a paraphrase of (1b), although (1b) will not be considered a paraphrase of (1a) in many circumstances.

(1)   a.   Susan broke the vase.
         b.   The vase broke because/when Susan dropped it.

In reality, it would be more fitting to ask how close in meaning two particular sentences are than whether they are paraphrases or not, because no reliable metric for such a binary decision is likely to be found.

Another source of complication is that acceptable ranges depend on the purpose of paraphrasing. In this respect, the recent study in machine translation (MT) by Resnik et al. (2010) is worth serious attention. They point out that available translation between world's languages is virtually limited to high quality-high cost by professional translators and low quality-low cost by MT. Using *Google Translate*, a free statistical MT service,[1] they propose a middle ground with the method they call *Targeted Paraphrasing*, which is expected to have potential to improve the quality of MT without human bilingual expertise (i.e., high cost).

In their approach, target-language monolinguals identify parts of an initial MT that are ungrammatical, nonsensical, or apparently incorrect. They then project such spans back to the source-text spans using word alignments. Source-language monolinguals are given the text and paraphrase the marked problematic parts. The solicited paraphrases are then passed through MT again. Resnik et al. report that their approach yielded substantial and consistent improvements in translation quality by all evaluation measures they employed. They anticipate that the human steps involved in their approach are also amenable to automation.

Two problems are found in Resnik et al.'s study. One concerns with the statistical MT method itself, whose parameters are derived from the analysis of huge bodies of bilingual (a minimum of one million words) as well as monolingual text corpora (a minimum of one billion words each). The statistical approach contrasts sharply with the rule-based and example-based MT methods (Och 2005). Presupposed in the statistical method is the existence of formal correspondences. If an expression in the source text becomes implicit in, or scattered over a long span of, the target text, the method is doomed to fail.

In this paper, we will discuss one salient example of this problem, i.e., the English words *less* and *fewer* when translated into Japanese, in which the concepts are neither lexicalized nor idiomatically-expressible. Because of this fact, Google Translate routinely fails conspicuously as discussed below.

We value Resnik et al.'s study because it opens the door to the possibility of human or mechanical manipulation to complement purely statistical MT. Based on this conviction, we

---

[1] See http://www.statmt.org/ for the statistical MT method.

point out the second problem of their research project. That is, they did not inquire as to which aspects of paraphrase led to the improvement of MT. What types of lexical or constructional material make the best candidate for improvements? What paraphrase techniques are most appropriate, and are there differences depending on the construction in question? Conversely, are there structures or lexical items that are resistant to paraphrase or which make MT more difficult even when paraphrased? Once such knowledge is accumulated, particular words or grammatical constructions in the source language could be automatically tagged for attention, so that a paraphrase is created either by a human or a machine before the material is subjected to MT.

A *Frames-and-Constructions Approach* to syntax and semantics inherently leads to predictions about what types of situation are amenable to paraphrase. This approach is based on the FrameNet project,[2] an online lexical database built on the principles of *Frame Semantics* (Fillmore 1982, 1985, 1994, Fillmore and Atkins 1992). FrameNet incorporates diverse types of linguistic information, more than any other online resources of a similar kind, and it is particularly utilitarian in investigation of paraphrase (Hasegawa et al. 2010, 2011, forthcoming).

In FrameNet, a *frame* is a schematic understanding of types of events, situations, individuals, and things, including the participants, props, parts, and their relations to each other and to the larger situation. Words are understood with respect to a particular frame as background. FrameNet groups words with the same background into frames, and defines these frames and the parts of the frame in prose. For instance, the words, *auction*, *retail*, *sell*, and *seller* are grouped in the `Commerce_sell` frame, for they all have as background a commercial transaction from the point of view of the seller of goods.

Each frame has a number of *frame elements* (FEs), which can be thought of as semantic roles. For instance, the `Attaching` frame involves at least three entities, an AGENT, an ITEM, and a GOAL; the AGENT causes the ITEM to be connected to the GOAL.

The English comparative construction, in broad strokes, has two variants: *equality* and *inequality*. These variant constructions are associated with frames and lexical/syntactic means of expression. The COMPARISON_EQUALITY construction uses *as* (sometimes *so*) to compare two entities or situations on some scale. For example, *We danced as much as them* quantifies the amount of dancing two groups did and indicates that the amounts are equivalent. The COMPARISON_INEQUALITY construction, on the other hand, is indicated by *more* (or equivalent morphology) and *less/fewer*. *We danced more than them* indicates that one amount is associated with a higher value than the other.

These constructions evoke frames of `Comparison_equality` and `Comparison_inequality`, with a common set of FEs: ENTITY, STANDARD, and FEATURE. The ENTITY is compared against some STANDARD (typically in an oblique expression, e.g., *than you*, *as last time*) with respect to their values for some FEATURE. The scope of the comparative morpheme is the FEATURE (*more <u>interesting</u>, <u>ran in marathons</u> less, <u>works</u> as hard*, etc.). The syntax and semantics of STANDARD in English can be quite complex, an issue which cannot be discussed in this paper.

The Frames-and-Constructions Approach to syntax and semantics predicts that these situations must be paid keen attention in translation when: (i) the target language lacks a construction corresponding to that of the source language; (ii) the target language has such a structure but it is dispreferred; (iii) the source language and target language have encoded a semantic domain with different frames; and (iv) those frames may be expressed in very different syntactic ways. In MT from English to Japanese, *less/fewer* crystalizes all of these issues.

---

[2] Information at: http://framenet.icsi.berkeley.edu/

Japanese has fairly productive, constructionalized means of expressing COMPARISON_EQUALITY:

(2)   **Standard-*to onaji-gurai* Feature**
             Kore-wa   are-to   onaji-gurai   omoi.
             this-TOP   that-with   same-about   heavy
             'This is as heavy as that.'

English allows straightforward negation of COMPARISON_EQUALITY, typically implicating less-than: *This isn't as interesting as that* (≈ *This is less interesting than that*). In Japanese, the same negated equality is expressed by a different set of morphemes:

(3)   **Standard-*hodo* Feature-*nai***
             Kore-wa   are-hodo   omoku-nai.
             this-TOP   that-extent   heavy-not
             'This is not as heavy as that.'

Japanese also readily conveys positive COMPARISON_INEQUALITY, i.e., English *more/-er*. Google Translate (as of 2012-07-20) translated (4a) perfectly as (4b).

(4)   a.   This laptop is heavier than that box.

        b.   *Kono   rapputoppu-wa   sono   hako-yori   omoi   desu.*
                this   laptop-TOP   that   box-than   heavy   COP.NPST

Less/fewer, however, is troublesome. As mentioned above, the notion of *less/fewer* is not lexicalized or constructionalized in Japanese, making straightforward translation difficult, if not impossible. Again, (5b) is a translation of (5a) by Google Translate.

(5)   a.   This laptop is less heavy than that box.

        b.   *Kono   rapputoppu-wa   sono   hako-yorimo   chīsai   omoi   desu.*
                this   laptop-TOP   that   box-than   small   heavy   COP.NPST
                (Incomprehensible)

(5b) is difficult to interpret. If forced, one might decipher it as *This laptop is smaller and heavier than that box*. That is, even if it is interpreted, it has a very different meaning from (5a).

     A close approximation of *less/fewer* is possible with negated COMPARISON_EQUALITY, as it conventionally conveys the intended meaning (in principle "not-equal" can mean "more or less than," but conventionally indicates "less than"). For example, a natural translation of (6a) is (6b). When fed with (6a), Google Translate produced an incomprehensible result, (6c).

(6)   a.   This is less interesting than that.

        b.   Natural translation
             Kore-wa   are-hodo   omoshiroku-nai   desu.
             this-TOP   that-as.much   interesting-not   COP.NPST
             'This is not as interesting as that.'

        c.   Google Translate

> *Kore-wa ika   omoshiroi   desu.*
> this-TOP below   interesting COP.NPST
> (Incomprehensible)

This failure is predictable, given that Japanese lacks a construction corresponding to *less/fewer*. If (6a) is paraphrased before going through MT as (7a), the quality of translation is amazingly improved, (7b), which is not only comprehensible, but also natural.

(7) a. This is not as interesting as that.

     b. Google Translate
> *Kore-wa   sore-to   onaji-kurai   kyōmi-bukai   mono   de-wa*
> this-TOP   that-with   same-about   interesting   thing   COP-TOP
>
> *arimasen.*
> is.not

     Lit. 'This is not a thing which is as interesting as that.'

The problem of the lack of linguistic correspondence is amplified by deep differences in expressions of quantification. English makes heavy use of nominal modification:

(8) a. More people participated than last year.
     b. Fewer animals were harmed than expected.
     c. Less than twenty liters of water remain in the container.

Japanese, by contrast, has strict limits on nominal quantification. It, instead, uses the adjectives *ōi* 'numerous/many' and *sukunai* 'scarce/few' predicatively, e.g., (9a).

(9) a. *Sanka-shita   hito-wa   sakunen-yori   ōkatta.*
        participated   people-TOP   last.year-than   were.numerous
        Lit. 'The number of people who participated was more numerous than last year.'

Here we can see a potential problem in Japanese-to-English translation. While the literal meaning of (8) can be conveyed grammatically in Japanese with nominal quantification, it is not always idiomatic or as natural as (8). Nevertheless, for *more*, nominal modification is possible with the phrase *ōku-no* 'numerous/many-GEN', as in (9b).

(9) b. *Sakunen-yori   ōku-no   hito-ga   sanka-shita.*
        last.year-than   numerous   people-NOM   participated.
        Lit. 'People more numerous than last year participated.'

For (8a), Google Translate returns:

(8) a. More people participated than last year.

(10)   *Yori   ōku-no   hitobito-ga   sakunen-yori   sanka-shimashita.*
      than   numerous   people-NOM   last.year-from   participated
      Lit. 'More people participated from last year.'

As shown in (9b), *sakunen-yori* 'than last year' must modify *ōku-no* 'numerous' (i.e., *more numerous than last year*). While *yori ōku-no hitobito* can be translated as 'more people', the STANDARD (*sakunen-yori*, 'than last year') is placed outside the subject NP in (10). This is permissible in English, but not in Japanese. In fact, *yori* occurs twice in (10) but can only be interpreted as 'than' once. Thus, the second *yori* must be interpreted as the homophonous postposition 'from', resulting in an incorrect translation.

When (8a) is paraphrased considering the Japanese strong preference of predicative quantification, vis-à-vis nominal quantification, the improvement is, again, substantial.

(8) a. More people participated than last year.

(11) a. The number of people who participated was more numerous than last year.

b. Google Translate
Sanka-shita   hito-no   kazu-wa   sakunen-yori-mo   ōku
participated   people-GEN   number-TOP   last.year-than   numerous

natta.
became

Lit. 'The number of people who participated became more numerous than last year.'

Nominal quantification with *less/fewer* is a different story. Some natural translations of (8b-c) are:

(8) b. Fewer animals were harmed than expected.
c. Less than twenty liters of water remain in the container.

(12) a. *Yosōshita-hodo   ōku-no   dōbutsu-wa higai-o   ukenakatta.*
expected-as.much   numerous   animal-TOP harm-ACC   did.not.receive
Lit. 'Not as many animals were harmed as was expected.'

b. *Yōki-ni-wa   mizu-ga   20-rittoru-mo   nokotte-inai.*
container-LOC-TOP   water-NOM   20-liters-as.much   remain-not.NPST
Lit. 'Not as much as 20 liters of water remain in the container.'

Running the sentences through Google Translate, we obtain incomprehensible results:

(13) a. *Yori   sukunai   dōbutsu-wa   yosō-ijō-ni   higai-o   ukemashita.*
more   scarce   animal-TOP   more.than.expected harm-ACC   received
(Incomprehensible)

b. *Miman-no   mizu   20-rittoru-ni-wa,   kontena-ni   nokotteiru.*
less.than-GEN   water   20-liters-LOC-TOP   container-LOC   remain
(Incomprehensible)

These examples demonstrate that it is necessary for translation systems and translation evaluation metrics to recognize that typological differences between languages may mean that the closest possible translation may result in radically different sentences in the source and target languages. In the case of negated COMPARISON_EQUALITY sentences, the difference is relatively minor. In the sentences above, however, the difference is significant, both lexically and syntactically.

In relatively well-defined areas of the lexicon and grammar, such as quantification, it is possible to set out a more-or-less definitive list of principles about which structures are preferred by which languages. Then, given a source sentence, one can generate a large range of typologically-attested grammatical or lexical patterns within the source language (even if some of the sentences are unnatural), and select the paraphrase whose structure best matches the typological profile of the target language for this domain. The end resulting translation, whether produced by MT or not, will likely improve.

We therefore support Resnik et al. in that paraphrasing has a tremendous potential for improving the quality of MT. Nevertheless, random paraphrases are unlikely to be effective when the cause of translation difficulty is rooted in profound typological differences. To further illustrate this point, we present the following translation by Google:

(14) a.   We need less fighting.

  b.   *Wareware-wa   ika-no   tatakai-o   hitsuyō-to   suru.*
       we-TOP   below   fighting-ACC necessity   do
       'We need the fights that are stated below.'

Many paraphrases, e.g. (15-16), still yield unacceptable translations by Google.

(15) a.   We need not to fight as much.

  b.   *Wareware-wa   onaji-kurai   tatakau   shinai-yō-ni-suru-o*
       we-TOP   same-about   fight   not-to-do-ACC

       hitsuyō-ga   arimasu.
       necessity-NOM   there-is

       (Incomprehensible)

(16) a.   We shouldn't fight this much.

  b.   *Wareware-wa,   kono   ōku-o   tatakau   beki-dewa-arimasen.*
       we-TOP   this   many-ACC   fight   ought-not-to
       'We shouldn't be involved in most of these fights.'

These translations exemplify that, contrary to Resnik et al.'s contention, certain problems in MT cannot be resolved by supplying paraphrase that is casually performed by native speakers who are unaware of the underlying issues. Knowledge of the typological possibilities of how a language might express a wide range of meanings is essential for at least some domains of translation tasks.

When simple, intuitive paraphrasing fails to provide better sentences for MT, a radically different approach must be taken. One might argue that because much of today's MT is statistical and incompatible with rule-based (symbolic) manipulation, which we suggest here. However, the issue we raise in this paper calls for special attention. MT is designed to translate pragmatic (vis-à-vis literary) texts, e.g., scientific, financial, political. In such genres, measurement and comparison expressions are shared needs of modern industrial society. Their meanings are relatively constant across languages and objectively describable. If statistical MT consistently fails in such a fundamental and essential field, some precautionary measures must be taken. At present, although we are uncertain whether or not purely statistical methods can eventually resolve this problem, it is clear that when a target language does not lexicalize a concept or when it lacks a construction in the source language, statistical analyses alone might not be able to identify corresponding patterns.

To briefly touch on possible future directions and more complex types of paraphrase, we can consider (17a) as a paraphrase of (14a).

(17) a. This much fighting is undesirable.

      b. *Kore-dake-no   tatakai-wa   nozomashiku   arimasen.*
        This-much-GEN fighting-TOP desirable   is.not
        (Unidiomatic but grammatical and comprehensible.)

Despite the semantic gulf between (14a), *We need less fighting*, and (17a), *This much fighting is undesirable*, they are relatable, and consequently derivable, in the Frames-and-Constructions Approach. In this case, the `Needing` frame, which is evoked in *We need less fighting*, is related to `Desirability` (a frame that includes such words as *desirable* and *undesirable*), and *less* and *much* are both instances of `Quantity`. As a pragmatic entailment of the `Quantity` frame, a negative comparative (*less*) is tantamount to a positive quantity (X *much*) plus the negation of the main predicate (the *un-* of *undesirable*). This chain of semantic reasoning, when combined with syntactic idiosyncrasies of the predicates involved, can lead us from *We need less fighting* to *We need less fighting*.

    This chain of reasoning is much more involved than the two comparison frames (`Comparison_equality` and `Comparison_inequality`) combined with negation, but we are confident that resources such as FrameNet (and a Frame-Semantic approach in general) can help reveal such paraphrase connections. This is done through a combination of frame relations and the syntactic and semantic differences between similar lexical items, much of which is recorded or derivable from FrameNet data.

    This paper has shown that there is a clear place for principled linguistic analysis in modern translation studies, even alongside statistical MT. Linguists have established many ways of analyzing the source of similarities and differences between such typologically divergent languages as English and Japanese. The Frames-and-Constructions Approach is especially powerful as it can recognize deep semantic correspondences and simultaneously handle syntactic and lexical differences. The tractability of such complex examples in this approach shows that there is a rich, untapped opportunity for augmenting MT with linguistic and typological analysis, especially when Frames and Constructions take center stage.

REFERENCES

Fillmore, Charles. 1982. Frame semantics. In *Linguistics in the Morning Calm*, ed. Linguistic Society of Korea, 111-137. Seoul: Hanshin.

Fillmore, Charles. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6: 222-254.

Fillmore, Charles. 1994. The hard road from verbs to nouns. In *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, eds. Mathew Chen and Ovid Tzeng, 105-129. Taipei: Pyramid Press.

Fillmore, Charles, and B.T.S. Atkins. 1992. Towards a frame-based organization of the lexicon: The semantics of RISK and its neighbors. In *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization*, eds. Adrienne Lehrer and Eva Kittay, 75-102. Hillsdale: Lawrence Erlbaum Associates.

Hasegawa, Yoko, Russell Lee-Goldman, Kyoko Ohara, Seiko Fujii, and Charles J. Fillmore. 2010. On expressing measurement and comparison in Japanese and English. In *Contrastive Construction Grammars*, ed. Hans C. Boas, 169-200. Amsterdam: John Benjamins.

Hasegawa, Yoko, Russell Lee-Goldman, Albert Kong, and Charles J. Fillmore. 2011. FrameNet as a resource for paraphrase research. *Constructions and Frames* vol. 3, no. 1, 104-127.

Hasegawa, Yoko, Russell Lee-Goldman, and Charles J. Fillmore. Forthcoming. In *The Socio-Pragmatics of Frame Semantics and Construction Grammar*, eds. Jan-Ola Ostman and Mirjam Fried. Amsterdam: John Benjamins.

Och, Franz Josef. 2005. Statistical machine translation: Foundations and recent advances. *The Tenth Machine Translation Summit*, Phuket, Thailand. http://www.mt-archive.info/MTS-2005-Och.pdf

Resnik, Philip, Olivia Buzek, Chang Hu, Yakov Kronrod, Benjamin Bederson. 2010. Improving translation via targeted paraphrasing. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 127–137. MIT, Massachusetts, USA, 9-11 October 2010. http://dl.acm.org/citation.cfm?id=1870671.