Predicting MT fidelity from noun-compound handling

John White

Litton PRC
1500 PRC Drive
McLean, Virginia
USA
white_john@prc.com

Monika Forner

Mendez, Inc.
5095 Murphy Canyon Road, Suite 300
San Diego, CA 92123
USA
Monika.Forner@lhsl.com

Abstract

Approaches to the automation of machine translation (MT) evaluation have attempted, or presumed, to connect some rapidly measurable phenomenon with general attributes of the MT output and/or system. In particular, measurements of the fluency of output are often asserted to be predictive of the usefulness of MT output in information-intensive, downstream tasks. The connections between the fluency ("intelligibility") of translation and its informational adequacy ("fidelity") are not actually straightforward. This paper discussed a small experiment in isolating a particular contrastive linguistic phenomena common to both French-English and Spanish-English pairs, and attempts to associate that behavior in machine and human translations with known fidelity properties of those translations. Our results show a definite correlative trend.

Keywords

Evaluation, fidelity, intelligibility, French, Spanish

Renewed efforts in MT evaluation have established a goal of determining ways of automatically evaluating MT output for the attributes relevant to particular tasks and stakeholders (Hirschman et al., 2000) Generally such approaches have involved capturing one phenomenon in the output that can be readily measured, and then attempting (or presuming) an extrapolation from these phenomena to the measurement of the output as a whole.

The connection between translation phenomena and the attributes of MT (e.g., fidelity, intelligibility, etc.) are not straightforward (Corston-Oliver, 2001; White, 2000; see also White, 2001, this workshop). In particular, it is presumed, but not demonstrated, that the apparent fluency of an MT output (measured, perhaps, by counting structural errors) will allow us to predict its usefulness in information-intensive tasks such as information extraction. Similarly, examination of the information conveyed in an output (perhaps through a variant of reading comprehension) may not directly predict suitability in a channel intensive task such as publication or broadcast. Whether holistic characterizations of MT output can be predicted by measures on a few phenomena remains unproven; yet any hope of rapid, ideally automatic MT evaluation depends on just such characterizations.

The current study arises out of a short project under the auspices of the 2nd MT Evaluation workshop, held in April 2001 in Geneva. This task is one of several which attempted to derive the attributes from known translation corpora by associating particular translation properties with other known characteristics of the texts in the corpus. In this study, we have examined the potential relationship between a particular linguistic phenomenon in the translated texts, and the known scores for fidelity for the same texts. The results of such a study, should there prove to be a correlation, will assist in the formulation of easily developed, possibly automated means of measuring one or a few phenomena, and predict the more general characteristics of the translation from those measurements.

Noun compounds in the DARPA MT Evaluation Corpus.

The body of translations known as the DARPA corpus consists of nearly two thousand translations from a variety of MT systems and expert human

translators. The corpus contains multiple translations of each of 300 source language newspaper articles, from French, Spanish, and Japanese into English. Additionally, each of these translations has been scored for attributes close to, or identical to, intelligibility and fidelity. Thus it is possible to isolate a particular phenomenon, measure it manually, and compare the results with the overall (already scored) attributes of the texts as a whole. By this means it may be possible to make the simpler, possibly automatic, measurement, and infer the entire attribute.

In a small experiment, we used a portion of the DARPA corpus to isolate and measure the behavior of noun compounds in both French-English and Spanish-English translations, and compared these with the pre-existing text scores for "adequacy" which is a measure of fidelity. The comparative syntax of noun compounds is similar for both French-English and Spanish-English pairs. Basically, expressions in English of the form N1 N2 must be expressed in French and Spanish as N2 *de* N1, that is, where N1 modifies N2. The reverse is not true: often, English can tolerate either order N1 N2 or N2 *of* N1. However, there are many occasions, even in relatively formal domains of discourse, where both options are not felicitous:

> *En Europa, bajaron las tasas de intereses* = In Europe, they lowered the interest rates ≠ …the rates of interests

Occasionally, the opposite of this case is true:

> *Hijos de Dios* = Children of God ≠ God Children

The issues of noun compound coverage are greatly exacerbated by interspersed modifiers, and the scope of modification in multiple-noun compounds:

> *le vice-ministre iranien de la défense* = the Iranian Vice-Minister of Defense ≠ *the vice-Iranian minister of the defense

> *la tête d'un consortium de quarante fabricants tchèques d'armement* = the head of a consortium of forty Czech arms manufacturers ≠ *the head of a forty consortium Czech armament manufacturers

Very often, the infelicitous structure may be grammatical, and may be in some semantic sense a translation of the source, but it is less intelligible, less fluent, than expected. Judgments such as these are, then, fundamentally of intelligibility. Thus in effect the comparison of noun-compound behavior to already scored fidelity behavior is a precise focus on a possible association between fidelity and intelligibility than that presented by White (2001, op. cit.).

Test Development.

The DARPA corpus contains approximately 600 translations from each of three language, including human translations (two for each text, used for control and reference) and the machine translations of a variety of systems representing a range of maturity. In 1994, literate, native English speakers followed a series of controlled instructions to rank, on a 1-5 scale, how successfully information contained in an expert translation was contained in a sample of machine translation (or human translated control) in an evaluation known as the adequacy measure (White 1995; Doyon et al. 1998).

Ranking these for their adequacy scores, we created a subset sample by selecting every 20th text in the adequacy ranking. In a preliminary examination during the 2001 Geneva workshop, we reduced the set further to the four lowest, four middle, and four best in adequacy scores from that subset, for each language. After examination of those results, we expanded to the every-20th set represented here. 33 texts were examined in each language pair.

In parallel to this procedure, we constructed a set of "theory-neutral" contrastive phenomena that obtain between the respective language pairs. For example, French and Spanish are more likely to place adjectives to the right of the nouns they modify; French and Spanish use a reflexive form to express passive more often than the be+passive participle common in English. In addition to the well-established grammatical comparisons between these languages and English, we added to the list by examining several translations to determine whether certain non-linguistic errors were common, for example the punctuation of numbers.

We selected handling of noun compounds for this study, after examining a set of translations and determining that this was a very common phenomenon in both the Spanish-English and French-English translations, and so capable of providing a good level of sample granularity.

For the selected texts in this subset sample, we observed the correct and incorrect noun compounds in the English output, referring to the source text at

the same time. We determined a score for each translation as the number of correct noun compounds over the total number. Certain constraints were imposed to optimize consistency across judgments, scorers, and language pairs. Notably, we scored for syntactic correctness rather, that is, scoring English noun compounds correct if their word order was, even when lexical choice, presence/absence of determiners, and other phenomena were not.

Results

Figures 1 and 2 show the results of the experiment. In Figure 1, the subset of French, scored by correct noun compounds, is arrayed against the adequacy rankings of the texts in the sample. Generally speaking, it appears that , the correctness of noun compounds increases with the adequacy scores. The same is true for the Spanish-English results in Figure 2. There are evidently several texts whose noun-compound behavior deviates from the pattern, but the low end of the range of these variations rises with the adequacy level.

The possibility of a consistent correlation between correct (syntactic) translation of noun compounds and MT fidelity is supported by these findings. It may be possible, then, to predict the overall fidelity of an MT system's output from the examination of a small subset of translation phenomena that occur. This finding, if ultimately valid, is significant particularly since the activity of determining correctness of noun compounds was, as we defined it earlier, something of an intelligibility exercise, and so the correlation suggests a relationship between

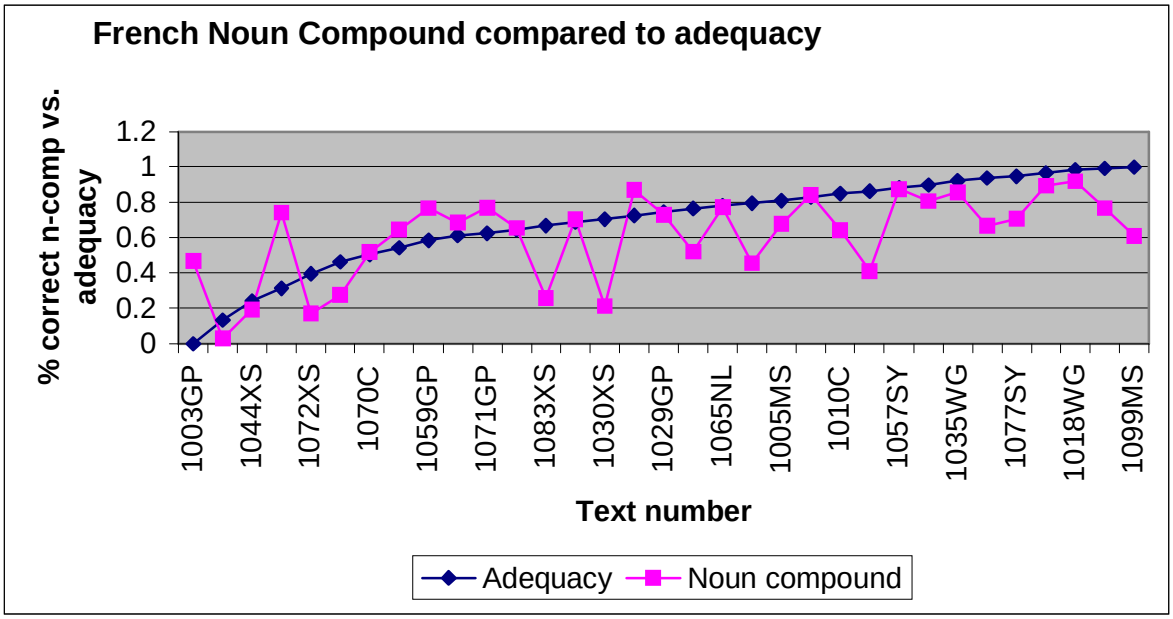

Figure 1: French noun compound correctness compared to text adequacy

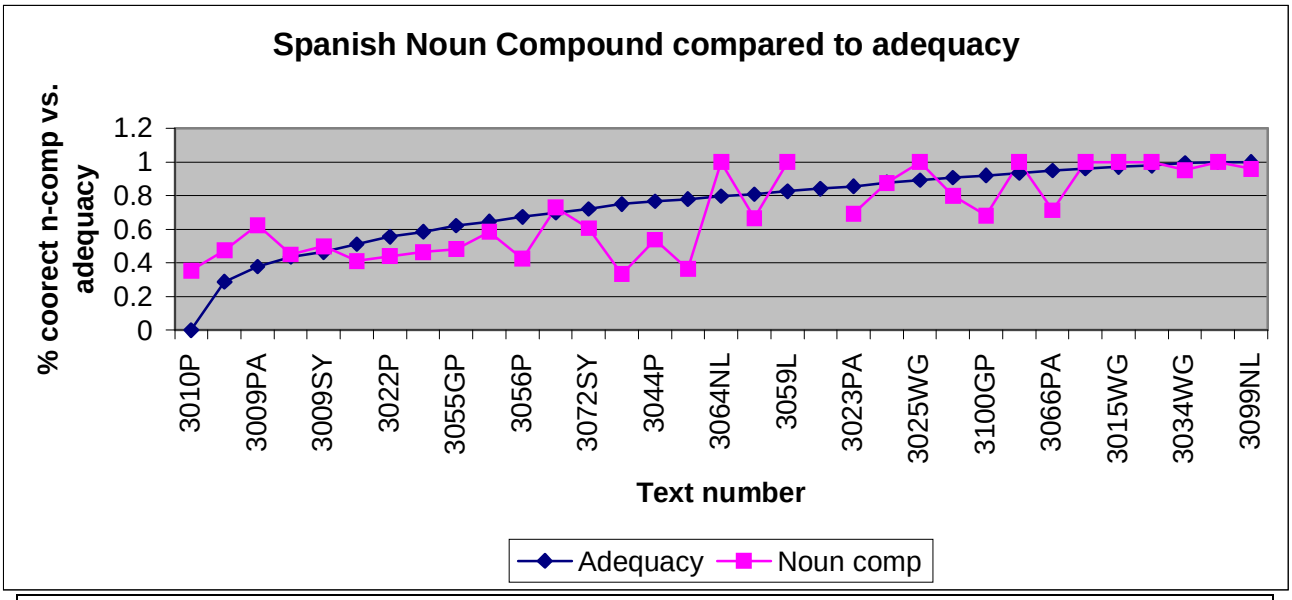

Figure 2: Spanish noun compound correctness compared to text adequacy

fluency and fidelity that can be exploited in number of ways toward automatic MT evaluation.

There remain several issues which should be addressed before implementing an approach similar to this one on a larger scale:

- The automatic identification of noun compounds, especially ill-formed ones, is quite difficult. In the short term, this study may serve best to verify that the correspondence between linguistic translation phenomena and information fidelity is evident and can be captured;
- Scoring of correctness of noun compounds, even given the constraints applied, remains subjective. All of the human subject factors come in to play in the experiment above, which would require a larger sample of scorers, and multiple views of the data, to control;
- We make some leaps from a correlation of noun-compounds and adequacy to a correlation between intelligibility and fidelity. It may be that the attributes measured by the DARPA series adequacy measure, while focused on fidelity, may not measure all aspects of it. As for the correlation to intelligibility suggested by the noun compound behavior, the French sample was also compared to the DARPA measure of "fluency" (an intelligibility measure) for the same texts, with results rather like those of the adequacy, namely, an upward trending, but imperfectly matching, pattern. This suggests that either noun compounds, the fluency measure, or both, fail to capture all aspects of MT intelligibility.
- As noted above, we attempted to score noun compounds on the basis of syntactic behavior, trying to ignore incorrect word choice. If there were a method of scoring that accommodates incorrect word choice, both the correlations of syntactic correctness of complex NPs with adequacy and/or fluency might turn out to be much higher. Moreover, it would be interesting to investigate to what extent incorrect word choice actually does or does not impede the possible effects of syntactic correctness on human judgments of adequacy and/or fluency.

Regardless of the ultimate usefulness of the specific observations reported here, it is vitally important to continue accumulate correspondences between translation phenomena and the attributes necessary for performance of MT in real-world information handling tasks. By doing more such studies we may ultimately transcend the temptation to presume, rather than demonstrate, the correspondence.

Bibliographical References

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A Machine Learning Approach to the Automatic Evaluation of Machine Translation. Proceedings of the 39th Annual Conference of the Association of Aomputational Linguistics. Toulouse, France.

Doyon, J., Taylor, K., & White, J. 1998. The DARPA Machine Translation Evaluation Methodology: Past and Present. *Proceedings of AMTA-98*. Philadelphia, PA.

Hirschman, L., Reeder, F., Burger, J., & Miller, K. 2000. Name Translation as a Machine Translation Evaluation Task. *Proceedings of the Workshop on Machine Translation Evaluation, LREC-2000*.

White, J. (1995). Approaches to Black-Box Machine Translation Evaluation. Proceedings of MT Summit 1995. Luxembourg.

White, J. (2000). Toward an Automated, Task-Based MT Evaluation Strategy. Proceedings of the Workshop on Evaluation, Language Resources and Evaluation Conference, LREC-2000. Athens, Greece.

White, J. (2001). Predicting intelligibility from fidelity in MT evaluation. MT Evaluation Workshop, 2001 MT Summit. Santiago del Compostela, Spain.