Pragmatics-based Translation and MT Evaluation

David Farwell, Stephen Helmreich
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003 USA
{david, shelmrei}@crl.nmsu.edu

Abstract

In this paper the authors wish to present a view of translation equivalence related to a pragmatics-based approach to machine translation. We will argue that current evaluation methods which assume that there is a predictable correspondence between language forms cannot adequately account for this view. We will then describe a method for objectively determining the relative equivalence of two texts. However, given the need for both an open world assumption and non-monotonic inferencing, such a method cannot be realistically implemented and therefore certain "classic" evaluation strategies will continue to be preferable as practical methods of evaluation.

1. Introduction

In this paper we present a view of translation equivalence derived from a pragmatics-based approach to machine translation. We argue that evaluation methods which assume a predictable correspondence between source and target language forms cannot adequately account for the data on which this approach is built. We describe a method for objectively determining the relative equivalence of two texts. However, given the nature of the requirements for this determination, we suggest that such a method cannot be realistically implemented and therefore certain "classic" evaluation strategies continue to be preferable. In what follows we will first sketch out a pragmatics-based approach to machine translation, define a notion of translation equivalence derived from such an approach, present a method for determining the degree of equivalence between to texts and, finally, discuss both the problems with this method and the implications of the underlying notion of equivalence for existing evaluation techniques.

2. Pragmatics-based MT

In this section, we outline a pragmatics-based approach to machine translation. Though this approach builds on concrete data provided by human translators and on hypotheses about the inferencing processes of the human translators, we do not make any claims for the psychological validity of this approach. In the next section, we use this approach to machine translation to develop a notion of equivalence that can be used in the evaluation of the results of any machine translation system, whether pragmatics-based or not.

Following a pragmatics-based approach, the process of translation is divided into two phases: interpreting the original source language text and constructing a target language text that conveys that interpretation. Both of these take place within a particular context which consists of models of the beliefs of the participants in the translation process and a model of the information conveyed during prior discourse. Because the beliefs of the participants are crucial to determining the content of the interpretation and the form of the target language text, and because beliefs vary from person to person, including from translator to translator, the approach entails that there are multiple legitimate translations for a given text which may vary significantly in terms of their meanings.

The process of interpretation involves constructing a structure of propositions which represents the information the author appears to have wanted to convey by a text (i.e., a reading). Since language underspecifies the information conveyed, much of the task involves filling out and connecting together the minimal content triggered by the actual text form. This is done against a background of information in the form of models of the beliefs of the speaker/author and of the addressee(s) and a model of the information conveyed thus far. The content of all of these models as well as of the interpretation itself are derived from a general ontology of world knowledge and an episodic knowledge base of particular objects, events and situations.

As an example, consider the following headline from a news article about a situation in Chile in which all the basic provisions in a city were being bought up by a nervous population in order to prepare for a possible pending earthquake. The article was translated from Spanish into English by two independent translators as part of the preparation of the DARPA machine translation evaluation corpus (White *et al.,* 1994). The original headline in Spanish reads:

Acumulación de viveres par anuncios sísmicos en Chile

The translations provided by the two translators are:

STOCKPILING OF PROVISIONS BECAUSE OF PREDICTED EARTHQUAKES IN CHILE

Hoarding Caused by Earthquake Predictions in Chile

These headlines reflect two different underlying interpretations of the source language text which the two translators arrived at. In the first case, the translator believes that the author is writing about government mismanagement, issuing an announcement to the press about possible pending earthquakes without adequately explaining the situation. The translator manifests this interpretation through lexical selection, stressing the likelihood and potential severity of the earthquake while downplaying the reaction of the population. In the above text segment, this is reflected in the translator's choice of *stockpiling of provisions,* a relatively rational activity in the face of an impending disaster, and *predicted earthquakes,* a relatively likely event, especially given modern scientific and technological capabilities. The second translator, on the other hand, believes that the author is writing about a case of journalistic sensationalism, blowing a story out of proportion to increase readership or listenership. This is manifested by stressing the overreaction of the population and downplaying both the likelihood and severity of any earthquake. Here that interpretation is reflected in the choice of *hoarding,* a selfish, antisocial behavior, and of *earthquake predictions,* focusing on the speculative aspect of the event.

Thus, each translator has brought to bear different knowledge in order to interpret the article and, as a result, each has arrived at a different interpretation and ultimately at a different translation. The initial Spanish expressions, *acumulación de viveres* and *anuncios sísmicos,* are neutral with respect to these interpretations as well as with respect to the eventual translations. To get to those interpretations and, consequently, to those translations, the translator must fill out the minimal information associated with the Spanish text and connect it in a coherent way with the rest of the interpretation. If one has as underlying assumptions that the government has dealt with the situation incompetently, that the press has reported the information professionally, that the earthquake is likely and the population has reacted predictably and appropriately, these gaps will be filled in and the connections established in quite a different way from one who has as underlying assumptions that the government has acted responsibly, that the press sensationalized the situation, that the earthquake is unlikely and that the population has overreacted. Thus it is that we assume that such underlying assumptions must be part of the interpretation as well.

In the second phase of the translation process, the task is to provide a target language form which expresses the interpretation (to the degree possible), that is, to provide a form which expresses the structure of propositions created to establish a coherent content and to connect it to the previously existing elements of the interpretation, along with the information used to produce that structure of propositions. This task is carried out in the context of the models of the beliefs of the translator and of the addressees of the translation along with a model of the information from prior discourse. The model of the translator will certainly be different from the model of the original author in that it will include many beliefs related to the target language community and its conventions that the original author would not have assumed. Similarly, the model of the addressees of the translation, as target language speakers, will be different from the addressees of the original document as well. Again, the content of all these models is derived from a general ontology of world knowledge and an episodic knowledge base of particular objects, events and situations.

Although the interpretations discussed above were described as part of the interpretation phase, it is possible that they could have been made during the realization phase at the point the translator is faced with actually selecting lexical items to express an element of the interpretation. Because languages differ in terms of the meanings of their expressions, connotatively as well as intensionally (or denotatively), the translator may be faced with various alternatives each having different implications. Returning to the example above, it is quite possible that the translators would not have been concerned about the specifics of the causal relation between the prediction and the amassing of goods until they actually considered the different implications of the different lexical choices English presents them with (i.e., *because of* or *caused by* among others). Whereas the former is somewhat vaguer in terms of the relationship referred to, easily admitting indirect causation or enablement as well as direct causation, the latter is more specific, essentially limited to direct causation. The translator, then, must consider which implications are appropriate for the situation. If the translator assumes a scenario in which the population is behaving appropriately given an impending catastrophe, then he or she might wish to play down the directness of the causal relation by selecting the former alternative. If, on the other hand, the translator assumes that the population has overreacted to sensationalist news reports, then he or she might wish to stress the directness causal connection by selecting the latter. In any case, this is what the translators did and it appears to be consistent with their respective views of the underlying intent of the author.

The process of selecting between lexical expressions, then, is essentially a process of deciding between which propositions (aspects of an expression's intension or connotation) may have to be added to or omitted from the interpretation in order to produce a fluent target language text. As in the case of interpretation, the desire to maintain coherence and connectedness is central and it involves reasoning on the basis of models of the beliefs of the translator and of the target language addressees and a model of the information previously introduced into the utterance context.

3. Equivalence

Given such an approach to the translation process, a natural approach to defining an equivalence relationship between texts (and this would apply to texts in the same language as well as between translations) is in terms of the degree to which their interpretations share the same set of readings or structures of propositions (i.e., have the same potential information content). At first blush one could actually design an algorithm to compare two structures of propositions with respect to their form and content. Expressions (or texts) in different languages which share a common set of readings or, at least, share a central reading would be more equivalent than expressions (or texts) in different language having disparate sets of readings or, perhaps, do not share any central reading. In fact, if no readings at all are shared, the two texts are simply not equivalent. (Equivalence, as we use it here, should not be understood as a mathematical equivalence relation. It is, rather, an indication of the percentage of readings shared.)

There are, in fact, two notions of equivalence under this general view. First, given fixed models of the author/translator and of the addressee of the text, construct all of the possible plausible readings, i.e., the set of possible, internally coherent structures of propositions, conveyed for each text to be compared. Each structure includes both the explicit and the implicit propositions which may be conveyed in that they are used to establish a coherent reading. For instance, the interpretation which eventually leads to the choice of *hoarding* in the translation for *acumulacion* requires adding the proposition that the behavior is antisocial to the structure of propositions representing that reading. This proposition would not be added to the reading of the second translation with *stockpiling of provisions* and, at least in that respect, the two translations do not share the same set of readings and are therefore not entirely equivalent to each other. However, both readings, in this case, are part of the set of possible readings that make up the interpretation of the source language text and so with respect to this particular proposition both translations are equally equivalent to the source language text. The former shares with the source language interpretation the reading containing the proposition; the latter shares with the source language interpretation the reading not containing the proposition.

The second notion of interpretation is broader. In this case, the models of the author/translator and addressees are not fixed but rather allowed to vary across all possible models that can be constructed given a fixed ontology and episodic knowledge base. In other words, rather than compare single interpretations (sets of plausible readings), the process is to compare sets of interpretations (i.e., sets of sets of plausible interpretations). We will not go into an example here although it might be worth pointing out that, at least with respect to the reading discussed above, the two translations will continue to have differing sets of interpretations and therefore not be fully equivalent to each other and yet each will be equally equivalent to the set of source language interpretations.

4. Evaluation

The pragmatics-based view of equivalence sketched out above opens up the theoretical possibility for a truly objective fidelity-based evaluation of translation quality. The approach with respect to establishing the fidelity of translation for the "core" readings is to fix the models of the source text author and the translator of the documents being evaluated and then proceed to provide all the plausible readings of these documents. The sets of plausible readings are then compared, reading by reading, to identify for each the similarity of the structures of propositions representing them. Variations between readings may be more or less severe depending on whether they include contradictory information *(hoarding* vs *stockpiling),* more fine-grained information (either *hoarding* or *stockpiling* vs *acumulación)* or more coarse-grained information *(acumulación* vs either *hoarding* or *stockpiling).* Deciding the whether two propositions are contradictory or the degree to which one might be a specification or a generalization of another would be based on ontological knowledge. The results of such an evaluation is a rating of equivalence of the source language and target language documents and, therefore, a rating of translation quality assuming that this content has been expressed in a fluent manner in each of the documents.

To establish the fidelity of a translation at the broader level of possible interpretations, the process described above must be repeated for every possible model of the source text author and of the translator which the ontology and an episodic knowledge base permit. It is assumed, of course, that the ontology and episodic memory used to construct all the other models or information structures is fixed. As a result of the iterated evaluation of the interpretations corresponding to each of the models, a series of equivalence ratings is produced which can be used to arrive at a global rating of equivalence. Various factors might potentially enter into calculating this global rating including weighting schemes based on the internal coherence of the readings making up the interpretation, on the typicality of the reading making up the interpretation and so on.

5. Problems and Implications

It should be immediately clear that while theoretically possible, the evaluation technique proposed above is impractical. Given an open-world assumption about the models of beliefs (even if the knowledge sources themselves are assumed to be closed), and the need for non-monotonic (defeasible) reasoning, it would be impossible to arrive at a representation of the possible interpretations of a text let alone compare such representations. It might be possible to restrict search to a relatively small number of possible limited interpretations if, for instance, a valid plausibility coefficient could be assigned to each reading and a valid coherence coefficient to each model of beliefs and only the most plausible readings were inspected for only most coherent sets of models. But as with so much that has been discussed here, these are very dubious assumptions. Furthermore, it is useful to keep in mind that translation does not take place in a vacuum, but rather always against some task requirement. This task may, of course, require some sort of general-purpose translation, but alternatively it might be related to question-answering, text summarization, information extraction, and so on. Each such application drives the interpretation of a text toward a different result. In essence, the task to be performed determines the appropriateness of the translation, as opposed to the semantics of the text, and so it is necessary to evaluate the quality of a translation in terms of the use to which the translation is to be put. A translation of engineering specs needs to be terminologically precise. A translation of the directions for putting something together needs to be simple, clear and sensitive to interpersonal conventions related to telling people how to do things. A translation of a document for information gathering purposes needs to be accurate with respect to its information content. A translation of an advertising brochure needs to be sensitive to and relate to the relevant sociological and psychological needs and predispositions of the target language community.

For instance, in translating a sentence such for, say, information dissemination purposes such as:

Pierre Vinken will join the board as a nonexecutive director.

it might not be necessary to interpret the expression *the board* as board of directors (as opposed to board of trade or school board, etc.) as all these might translate into the target language in the same way. It is certainly unlikely that given such an end use a translation system would have to resolve the implicit reference to the particular company whose board is being mentioned.

But in the context of an information extraction task, where the goal, say, is to identify all the reported changes in corporate boards for some period of time, it is likely that the translation would have to be modified to include such information. The problem is that, even when considered within its context, it is unclear which company's board Mr. Vinken is going to join. Plausible interpretations include the company of which he is already executive director, i.e., Elsevier NV, a new corporate group that is being formed of which Elsevier is one parent, Reed Elsevier, the other corporate member of this group, Reed International PLC, or some other company altogether. Assuming that different translators will legitimately arrive at different interpretations, there might be up to four different translations associated with the sentence. In any case, the end use of information extraction drives the search for a relevant company while the information in the discourse context provides that possible alternatives and, together with a general knowledge of the world, supports the inferencing process.

In response to what we have claimed, it might be argued that what is going on here is that different grain sizes of a translation are relevant for different tasks (i.e., for information retrieval, for example, one might need only lexical equivalents without any corresponding text structure at all). If so, then perhaps all possible tasks could be accommodated by using a translation that is, in some way, "maximal."

We would counter, however, that such a maximal translation must be of either of two types. On the one hand it might be a translation based on an exhaustive interpretation of the input text. That is, it would convey not just the semantics of the expressions in the source text, but also all of the inferences that could be drawn on the basis of the world knowledge and the models of beliefs of the author and translator and of the addressees. We argue that such a translation would almost certainly differ from one translator to the next, since the background beliefs of each translator will vary, as will the selection process by which relevant beliefs are brought forward for inferencing. In addition, we have already implied in the discussion above concerning an object evaluation method based on equivalence that such an exhaustive translation would be impractical if not impossible to produce.

On the other hand, it might be a translation of the only the direct and directly-inferable information expressed by way of the text. That is to say, any pragmatic inferencing based on world knowledge would be postponed until the application stage as needed by the task at hand. However, we claim that such a translation would be inadequate for such inferencing at that point because the inferencing mechanism would no longer have access to the assumed world knowledge of the source text author, nor does it have access to the actual text itself.

What can be said about standard form-based evaluation methodologies, e.g., word error rate-based methodologies such as described in (Och and Tillmann, 1999), methodologies based on syntactic equivalences such as described in (Lehmann *et al.,* 1996), and so on, is that they are not sensitive to the potential variation of legitimate translations of a text. Even semantic-based comparisons, such as described, for instance, in (Rajman and Hartley, 2001) are likely to be overly restrictive. Our investigation of the DARPA MT evaluation corpus (White *et al.,* 1994), which consists of two conservative (or "literal") translations of a given documents into the same target language by different translators, shows that up to 40% of the corresponding (constituent-level) translation units in two translations differ in some respect. Up to 60% vary if a third independent translation is added to the mix. Almost 20% of the translation units show significant variation in that they express differences in the information content conveyed. For example, as we have discussed elsewhere (Farwell and Helmreich, 1999), with respect to one of the evaluation corpus text sets, the Spanish expression *tercer* was rendered as *third* by one translator but as *fourth* by the other while *segundo* was rendered as *second* and *third* respectively. Extensionally, the two translations may be equivalent but semantically they are clearly not. Thus, one should not be surprised to find significantly different translations of the same input text.

This observation is in some sense tacitly supported by the development of the Bleu Machine Translation evaluation methodology at IBM (Papineni *et al.,* 2001). Because prior measures based on word error rates, which had been used successfully for speech recognition evaluation, were overly restrictive when applied to machine translation, a new approach has been developed which permits having multiple different target language translations against which to calculate the error rates. In essence, the shift from a single target translation to multiple target translations is motivated by the potential for variation among legitimate translations. The problems are that there are hundreds of such possible variants of a typical news article, not merely 5 or 10.

Equally problematical are evaluation techniques based on comprehension exercises such as multiple-choice question based evaluations, as for example (Leavitt, *et al.,* 1971) and (Orr and Small, 1967) or the knowledge test technique proposed in (Sinaiko, 1979). For instance, in the earthquake article described at the outset of the paper there were various possible questions that might have been asked that would be answered differently depending on which translation one read. Examples include:

- What is the expected size of the accompanying tidal wave?
- How long has it been since the last earthquake?
- How likely is it that an earthquake will take place?
- How would characterize the reaction of the population?
- How would you characterize the actions of the government office for emergencies?
- How would you characterize the objectivity of the press?

The answers to these questions will be different according to the translation one reads. In one case the wave will be as high 20 feet, the last earthquake was over 100 years ago, the population behaved appropriately given impending disaster, the office of emergencies was incompetent and the press professional. In the other case, the wave might reach as high as 20 feet, the last earthquake was 100 years ago, the population panicked, and the office of emergencies acted appropriately but the press sensationalized the predicted earthquakes.

In conclusion, it appears to us, at least, that the classic methods involving human bilingual subjective evaluation for fidelity and monolingual subjective evaluation for naturalness of expression continue to be best techniques for evaluating MT quality, such as (Halliday and Briss, 1977) or (Crook and Bishop, 1965). Although they may be more expensive and may only be feasibly applied to a small percentage of the total translation corpus, they are nonetheless the only method which can deal effortlessly with potential legitimate variations based on differing interpretations or driven by different end applications or uses.

6. References

Crook, M., and H. Bishop. 1965. *Evaluation of Machine Translation.* Final report, Institute for Psychological Research, Tufts University, April 1965.

Farwell, D., and S. Helmreich. 1999. Pragmatics and Translation. *Procesamiento de Lenguaje Natural,* 24: 19-36.

Halliday, T., and E. Briss. 1977. The Evaluation and Systems Analysis of the Systran Machine Translation System. *Report RADC-TR-76-399,* January, 1977. Rome Air Development Center, Griffiss Air Force base, New York.

Leavitt, A., J. Gates, and S. Shannon. 1971. Machine Translation Quality and Production Process Evaluation. *Report RADC-TR-71-206,* October 1971. Rome Air Development Center, Griffiss Air Force Base, New York.

Lehmann, S., S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP — Test Suites for Natural Language Processing. *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-96).

Och, F., C. Tillmann, and H. Ney. 1999. Improved

Alignment Models for Statistical Machine Translation. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora.*

Orr, D., and V. Small. 1967. Comprehensibility of Machine-Aided Translations of Russian Scientific Documents. *Mechanical Translation and Computational Linguistics,* 10, 1-10.

Papineni, K., S. Roukos, T. Ward, W-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report RC22176(W0109-022),* September 17, 2001.

Rajman, M., and A. Hartley. 2001. Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. *Proceeding of the MT Summit VIII Evaluation Workshop: Who did What to Whom,* p. 29-34.

Sinaiko, H. 1979. Measurement of Usefulness by performance test. In G. Van Slype (ed.), *Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management.* Report BR-19142. Bureau Marcel van Dijk.

White, J., T. O'Connell, and F. O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas,* 193-205.