IN DEFENSE OF RATIONALIST APPROACHES TO MT RESEARCH

David L. Farwell
Computing Research Laboratory
Las Cruces, NM 88001
tel: 505 646 5108
email: david@nmsu.edu

Abstract:[1] This paper presents a weak defense of rationalist approaches to Machine Translation (MT) based on the following line of reasoning. There is, independent of non-linguistic context, a one to many relationship between a source language text and the possible high quality target language translations of the text. This is due to the fact that there is a one to many relationship between a source language text and its possible interpretations against different non-linguistic context and between an interpretation and the possible target language forms that can be used to express that interpretation which vary according to non-linguistic context. To model such relationships requires the investigation of a source language text along side of a range of possible high quality translations of that text Machine translation systems can be (weak), if not must be (strong), developed on the basis of such models.

1. Introduction

1.1 Goal

In this paper, I wish to address the issue put to the workshop contributors: rationalist vs empiricist methods in MT—What is the single, most important source of the knowledge that MT developers must build into their systems? Is it the theories and intuitions of the MT researchers themselves? Or, is it large, bilingual corpora in which this knowledge is already latent, waiting to be extracted by various automated procedures?

I have several problems with the question as posed so I will begin by redefining it in a form that I understand and then present an argument that a rationalist methodology presents a reasonable approach to the MT task. The argument is not that it is the only viable approach because it is difficult to imagine what would constitute a knock-down case for or against either methodology.

1.2 Clarification of the Issue

The ultimate goal of Machine Translation is to build a device that does something, namely, translate documents in one language into documents in another language. The target language document should provide the reader with equivalent information for equivalent purposes (unless, a priori, we know

[1] I wish to thank Stephen Helmreich, Yorick Wilks, Eduard Hovy, Sergei Nirenburg and the anonymous conference reviewers of the this paper for their comments and suggestions. They have greatly helped to clarify the discussion.

that the intended use of the target language document is not the same) in an appropriate (natural, readable, well-formed, etc.) target language form. But if MT systems are the product of development, then formulating models of MT systems and/or how to build such systems is the central objective. If elaborating models of MT systems and/or how to build such systems is the objective of MT research, then the question is whether an empiricist or a rationalist methods for constructing such models and will lead to better results, that is, better models. The problem is that "better" requires an evaluation metric and those involved in MT have, appropriately, identified numerous such metrics but not agreed as a group to any.

What is not at issue is whether one or the other approach is empirical or unempirical (or anti-empirical) or whether empirical versus unempirical approaches are preferable. No one involved in MT has the luxury of pursuing theological approaches based on beliefs and intuitions. Everyone is involved in MT is involved in empirical investigations. Similarly, not at issue is whether one or the other approach leads to solid engineering or to solid science. The issue is irrelevant if we are discussing methods for improving system designs of and/or strategies for constructing a prototypical system removed from a specific application. If we are not discussing that, then we need to face a slew of issues regarding the definition of the specific task, of the specific users, of existing and available software, of existing and available expertise, and so on where methodology becomes a moot point.

For both approaches, a general methodology is to hypothesize, test, and revise. Data plays a key role in both approaches since hypotheses are tested against data, and succeed or fail according to how they account for (predict, describe, explain) data. Hypotheses play a key role since they fail informatively (suggest the next problem to be resolved and perhaps even how to resolve it) or uninformatively (suggest nothing). The differences between the two approaches appear to me to lie in the answers each provides to two questions.

- Where does the data come from?
- Where do the hypotheses come from?

For the first question, the empiricists answer the world, which is full of naturally occurring examples of translation. The rationalists answers the laboratory, where the data can be collected under controlled conditions for the purpose of setting up valid tests of a given hypothesis. For the second question, the empiricist answers they "emerge" from the data. That is, given any explicit, formal, (rational) mathematical model (whether that is selected arbitrarily or by rational choice, say, on the basis of an analogy between the type of the data to be covered and the type of data to which the model has been successfully applied in the past), the hypotheses result from applying the model to the data. The rationalist answers it does not matter where the hypothesis comes from. Anywhere will do, including the researcher's intuition although it is more likely to be the researcher's ability to formulate generalizations from inspecting data, so long as it are explicitly formulated and testable.

2. The Argument

The argument to be presented runs as follows. Texts in any language are vague in that there is a one to many mapping between the text and the possible coherent interpretations of that text independent of a non-linguistic context. Similarly, given a coherent interpretation, there is a one to many mapping between the interpretation and the possible texts that might be used to express it independent of a non-linguistic context. Thus, for any given text there are many potential high quality translations from the point of view of surface text to surface text correspondence independent of a non-linguistic context. The range of variation is reflected by the fact that the interpretation of a given expression in a given context by one person is not necessarily the same as that for another person or, obviously, in another context. Linguistic expressions require lexical and structural disambiguation, the resolution of anaphoric and cataphoric references, the recovery of ellipted information and metonymic references, the interpretation of metaphors and so on. To arrive at a sufficiently complete interpretation, the addressee must infer a great deal to make sense of the utterance. Since different contexts and different addressees with different assumptions about the context and the world in general set up different inferences, interpretations vary.

A rationalist account of each potential variant which takes into account the non-linguistic context provides the possibility of dealing with this one to many relationship. Were it possible to show that empiricist methodologies are too weak to deal with the extensive potential variation of equivalent texts, the argument could be used against such approaches. This, however, appears to me to be a priori impossible. If there is a rationalist account of each variant than there must be some empiricist account. Thus the argument is a weak one.

To make a case then, I need to show that variation in high quality translation is potentially extreme at the level of surface text and that a rationalist approach, based on modeling the context, speakers, addressees, is, at least, tenable. To do this, I will present and discuss example data that reflect the problem. The discussion focuses on the data and some hypotheses which hint that the answers to the two method distinguishing questions posed above are at least compatible with a rationalist methodology.

2.1 Data

The following is a news article taken from Reuter's News Service. It provides a context for interpreting the example sentence.

> - By Patricia Zengerle PITTSBURGH, April 26, Reuter -
>
> American Telephone and Telegraph Co's fight to acquire NCR Corp remained unresolved Friday afternoon, even though five days have passed since what seemed a breakthrough in negotiations. "Nothing's happened that I know of. It is all quiet," said a spokesman for AT&T. He said he also had no information on whether discussions had continued during the

> week. An NCR spokesman also said he was not aware of any developments. On April 23 it asked AT&T to guarantee a 110 dollar per share price to stockholders even if AT&T's shares drop to 32-1/2. AT&T is offering a guarantee to 35-1/2. A source close to the company said AT&T had rejected NCR's request for the guarantee to a 32-1/2 share price in the course of negotiations before AT&T announced its 110 dollar per share bid on April 21. Though the source said he knew of no contact between the two sides Friday, there had been some between the companies' bankers during the week. AT&T was up 3/8 to 38 Friday. NCR was up 3/8 to 103-7/8. Analysts said they did not consider the delay a cause for concern about whether AT&T would succeed in buying NCR, though AT&T's 110 dollar bid led many to predict a deal within days. NCR had said previously that it would consider the merger at that price. After AT&T announced the higher bid, NCR Chairman Charles Exley said he was prepared to recommend the merger to NCR's board if AT&T would provide a guarantee of 110 dollars a share in the stock for stock transaction. "The whole process has taken longer than expected," said Jay Stevens of Dean Witter Reynolds. AT&T's four nominees to NCR's board will become NCR directors May 1. NCR said it had not decided whether its board would meet before then.

The text was presented to a number of bilinguals[2] and they were asked to translate the last sentence (1).

(1) *NCR said it had not decided whether its board would meet before then.*

This is classic rationalist methodology. The collection of data is controlled since we need multiple examples of the same translation, something which is very rare in the real world, in order to investigate what variations are possible. Despite the fact that the data was not naturally occurring in the empiricist's sense, however, it is not artificial data and a more serious version of the methodology could certainly be carried out, I assume, with the same general results. Here is a sample of the results:

SPANISH

(2) a. *NCR dijo que no ha decidido si su comité podría reunirse antes.*
NCR said that not has-it decided if its board would-be-able-to meet earlier

b. *NCR dijo que no se ha decidido si su comité ejecutivo se juntaría antes de esto.*
NCR said that not self has-it decided if its board of-directors self would-meet earlier than this

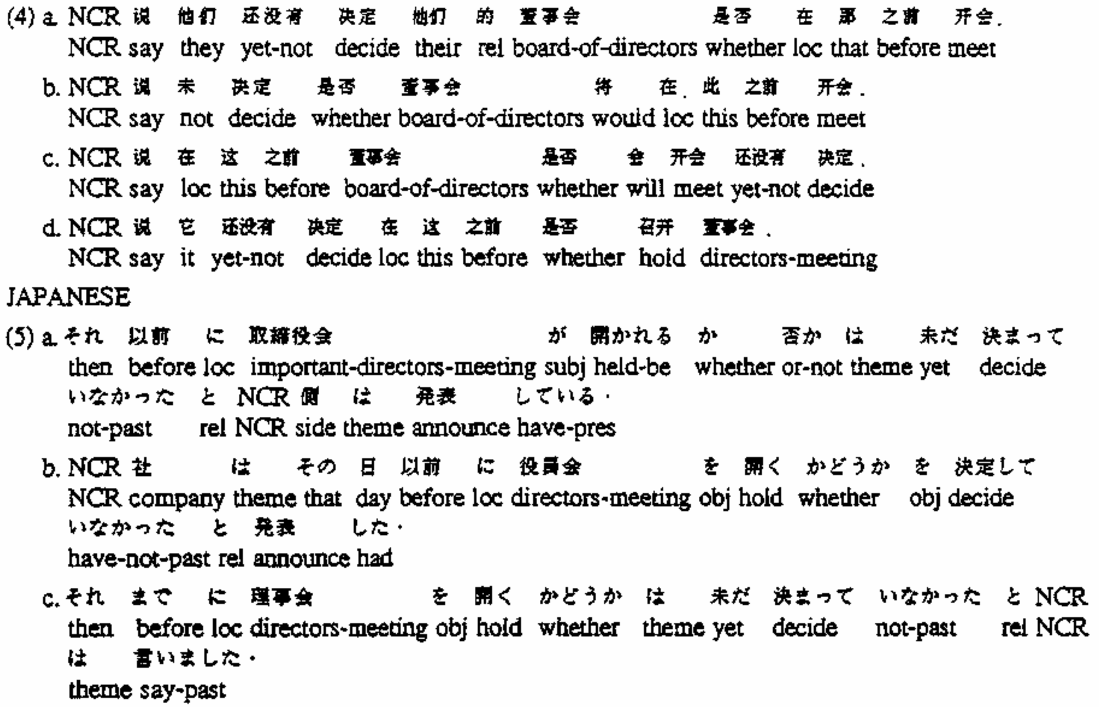
GERMAN

(3) a. *Die Gesellschaft NCR hat veröffentlicht, dass sie noch nicht entschieden hat, ob ihre*
the company NCR has announced that it yet not decided has whether its
Direktoren sich vorher treffen würde.
directors self beforehand meet would

b. *NCR sagte dass es nicht entschieden wurde ob das Direktorium sich eher versammeln*
NCR said that it not decided was whether the board-of-directors self earlier meet
wüurde.
would

[2] These included both first language English speakers and first language Spanish, Japanese, Chinese and German speakers. Some were professional translators but most were not. There is no pretense here to well-controlled methods of data collection since the point is merely to illustrate the problem, namely, the underdetermination of information in an interpretation.

CHINESE



(4) a. NCR 说 他们 还没有 决定 他们 的 董事会 是否 在 那 之前 开会.
NCR say they yet-not decide their rel board-of-directors whether loc that before meet

b. NCR 说 未 决定 是否 董事会 将 在. 此 之前 开会.
NCR say not decide whether board-of-directors would loc this before meet

c. NCR 说 在 这 之前 董事会 是否 会 开会 还没有 决定.
NCR say loc this before board-of-directors whether will meet yet-not decide

d. NCR 说 它 还没有 决定 在 这 之前 是否 召开 董事会.
NCR say it yet-not decide loc this before whether hold directors-meeting

JAPANESE

(5) a. それ 以前 に 取締役会 が 開かれる か 否か は 未だ 決まって
then before loc important-directors-meeting subj held-be whether or-not theme yet decide
いなかった と NCR 側 は 発表 している.
not-past rel NCR side theme announce have-pres

b. NCR 社 は その 日 以前 に 役員会 を 開く かどうか を 決定して
NCR company theme that day before loc directors-meeting obj hold whether obj decide
いなかった と 発表 した.
have-not-past rel announce had

c. それ まで に 理事会 を 開く かどうか は 未だ 決まって いなかった と NCR
then before loc directors-meeting obj hold whether theme yet decide not-past rel NCR
は 言いました.
theme say-past

The quality of the translations can be debated, as is always the case, even when very experienced, highly trained translators are involved. Some of the translations are arguably wrong with respect to one or another part of the translation. Nonetheless, the translations represent adequate, certainly accurate, translations.

2.2 Discussion

The example sentence includes cases of metonymy, anaphora, ellipsis, and lexical and syntactic ambiguity, ail of which are useful for demonstrating the vagueness of linguistic expression. The discussion will focus on the first three phenomenon, showing that they are not pseudo problems for machine translation but rather central to the development of any MT system.

The use *of NCR,* which strictly speaking is used to refer to a legal entity, to refer metonymically to agent of *said* and, indirectly through anaphoric reference, to the agent of *had not decided,* requires the translator to make inferences. The expression underspecifies the potential referents and the translator, either by knowing specifically who or what the author of the text was referring to or on the basis of a general knowledge of what "sayings" or "decidings" in this context involve, must construct a model of the situation being described. In particular, the translator must hypothesize reasonable agents of the "saying" and the "deciding" being reported. In the first case, it would be reasonable to hypothesize a spokesperson for the company, perhaps the same spokesperson mentioned explicitly in the fourth sentence of the paragraph. This is not the only possibility. It could have been an unnamed source close to the company or it might have been the CEO of the company mentioned in the 13th sentence or any of numerous other people. In the second case, where *NCR* is used via anaphoric reference to refer to the agent of *had not decided,* the translator must infer who (or what) could make such decisions. A reasonable choice here might be to hypothesize something such as "the relevant decision makers at NCR" but, again, the CEO or some executive committee or any of a number of other possibilities might be inferred. The point is that in one form or another, the translator has to resolve the reference through inference based on the information in the context (in the broad sense).

Having inferred referents, the translator then has to decide how to make a corresponding reference using the target language. This decision is constrained by the requirements of the target language grammar, the need to be explicit without being overly explicit, stylistic requirements given the genre, and, of course, the desire to be provide as near as an equivalent as possible given the intended use of the translation. In the case of the agent of the "saying", most of the translations have resorted to a corresponding metonymic reference (Spanish 2a and 2b, German 3b, Chinese 4a, 4b, 4c, and 4d, and Japanese 5c). This is interesting in that metonymy is by no means equivalently productive across languages although in this case, apparently, there is no special problem. More interesting in terms of the discussion, however, are the translations where alternative expressions have been used, that is, the equivalents of "the NCR company" (German 3a and Japanese 5b) or of "the NCR side" (Japanese 5a). While these expressions are also used metonymically to refer to people, they nonetheless indicate that the reference of CR has been resolved since the metonymic use of similar expressions which refer to things other than companies might not be possible.

What happens with respect to the treatment of the agent of the "deciding" is also informative. In this case, an explicit equivalent of *it* appears in only three examples (Spanish 2a and German 3a). The most common strategy, rather, appears to be to make no reference to an agent of the "deciding" at all, thus avoiding a commitment to identifying exactly who (or what) is doing the "deciding". This is done either by passivizing the predicate, changing *had not decided* to "has not been decided" (Spanish 2b, German 3b and Chinese 4d) or by omitting a subject expression in those languages that permit that option (Chinese 4b and 4c and Japanese 5a, 5b, and 5c). Perhaps most illuminating in terms of motivating the need for rilling out an underspecified interpretation is the Chinese translation in (4a) where the equivalent of "they" is used. It must be the case that the translator at least assumes the "deciding" is made by a group of people at NCR. Thus with respect to metonymy, we have attempted to show that the <u>data</u> reflects a wide range of possibilities, that an account of each the possibilities is plausible if we pursue a rationalist methodology in which we have recourse to all of the relevant information for inferring a coherent interpretation. We have not, of course, shown that the possibilities are sufficiently numerous, or from the point of view of a surface text to text correspondence, unpredictable, to cast serious doubt on an empiricist methodology.

We now turn to the question of interpreting the reference of anaphors, here *it,* the subject of *had not decided, its,* the possessive determiner of *board,* and *then,* the object of *before.* We have discussed the first case and found that in most of the translations, the *it* of the source language text has no equivalent since the translators chose to passivize the predicate (Spanish 2b, German 3b and Chinese 4d) or simply to omit reference to it (Chinese 4b and 4c and Japanese 5a, 5b, and 5c). In one case (Chinese 3a), the equivalent of "they" is used as the equivalent indicating the the translator at least assumed the decision is to be made by a group of people. In only three cases (Spanish 2a and German 3a) was the equivalent of "it" used. What all of these examples share, however, is the need to resolve the referent of it through "NCR". This is by no means necessarily the case but rather depends on what makes sense given a very underspecified reference to the agent of the "deciding". What makes sense is not a question of local formal constraints of English (i.e., resolve the reference to the preceding nominal) but rather of the what, in the broader context of the article, is being described.

As in the case of the subject of *had not decided,* the identification of which *board,* expressed as the possessive determiner *its* in the source language sentence, is in most cases simply omitted (Chinese 4b, 4c, and 4d and Japanese 5a, 5b, and 5c). Again, the equivalent of "their" is used in Chinese, (5a), implying that the translator is interpreting *its* as referring via *it* via *NCR* to a group of people. The equivalent of "the" is used in the German translation in (3b) which is tantamount to the omission strategy in Chinese and Japanese since there is no indication of a possessor. Only in the translations in Spanish (2a and 2b) and German (3a) is an equivalent possessive determiner used. Again, common in all cases is the resolution of the anaphoric reference ultimately through *NCR* and, again, it is not surface textual constraints that guide resolution but rather making sense of what is being described given the information in the broader context What accounts for the differences is the translations across languages are, on the one hand, the alternative (rationally justifiable) interpretations, and on the other, the expressive possibilities of the different languages given the need to state as succinctly and yet as informatively as possible the situation described.

The range of forms used to express *then* include nothing in those cases where the translator chose to use an adverbial such as "earlier" (Spanish 2a and German 3b) or "beforehand" (German 3a) as an equivalent of *before then,* "this" (Spanish 2b and Chinese 4b, 4c and 4d), "that" (Chinese 4a and Japanese 5a and 5c), and "that day" (Japanese 4b). While most of these expressions could be selected without resolving the reference of the pronoun, the Japanese translation in (4b) is only possible if the translator realizes that "then" is used to refer to May 5th. In this case, the range of possible referents is simply not within the local context.

The third case concerns the interpretation of the ellipted information. The first instance involves the recovering the missing "of directors" associated with *board.* Note that this must have occurred in the case of the translations in Spanish 2b, German 3a, 3b, Chinese 4a, 4b, 4c, 4d, and Japanese 5a, 5b and 5c. In each translation, the equivalent of "the directors" (rather than "board") is used. Indeed, it is only in the Spanish translation in (2a) that the translator has used an equivalent of "board" per se to make the reference. It is important to note that there are other types of "boards" (e.g. of education, transportation, governors, etc.) so that it must be the case that these references to "directors" come from making sense of *board* in this particular contexts based mainly on the mention of the board of directors of NCR in the prior sentence.

The second case of ellipsis concerns the insertion of the equivalent for "yet" in the expression corresponding to *had not decided* (German 3a, Chinese 4a, 4b and 4d, and Japanese 5a and 5c). Here, whether due to grammatical limitations or to stylistic convention, the translators had to insert some adverb meaning roughly "up to that point"- which is implicit in the English source language text. It is not the only type of temporal adverb that could be inserted formally although, in this context, it is clearly makes sense. The point is that, again, it is necessary to fill out the underspecified properties of the situation. It difficult to see how, based solely on surface text, it is possible to identify just those contexts where this insertion is natural.

3. Conclusion

The results of all this is that a plausible minimal interpretation of (1) would be something like:

> *A spokesperson for NCR reported that the relevant decision makers at the company had not yet made the decision as to whether the company's board of directors would convene prior to that date [May 5].*

This interpretation is but one of perhaps countlessly many interpretations of (1) that make sense in the broader context. It says more than (1), of course, and is minimal in that it is superficially coherent. But there are many possible events which are compatible with the interpretation provided and the trick for translation is to use the target language to set up a scenario in the addressee's mind that is at least as restrictive as the interpretation provided without being overly explicit, without violating any of the grammatical or semantic conventions of the language, without violating the stylistic conventions of the language given the genre and, to the extent possible, being only as informative as the source language author.

To address the problems described, a rationalist would argue that researchers must understand how translation is done by formulating explicit hypotheses, testing them against data and identifying how those hypotheses failed. This appears to be at least a reasonable way to proceed. The question for the empiricist is whether, given the data, it is plausible to attempt to induce, from the inspection of data, hypotheses (the nature of which are unimportant by the way so long as they are explicit and testable, e.g., they need not be plausible explanations from a psychological point of view or, to put it the other way round, they certainly could be information theoretic) that account for that data and then new data.

In closing, it should be noted that both traditions have a long history of repeated failure in research on language if success is measured as producing fully automated high quality translation. If success is measured in other ways, both have histories of success. For many, then, the question is not what is the single, most important source of the knowledge that MT developers must build into their systems but rather which type of research leads to useful and interesting models.