MT 2010 -- Towards a Road Map for MT
Santiago de Compostela (Spain), 18 September 2001

Towards Pragmatics-based Machine Translation

David Farwell, Stephen Helmreich

Computing Research Laboratory
New Mexico State University
Las Cruces, New Mexico
USA
david/shelmrei@crl.nmsu.edu

Abstract

We propose a program of research which has as its goal establishing a framework and methodology for investigating the pragmatic aspects of the translation process and implementing a computational platform for carrying out systematic experiments on the pragmatics of translation. The program has four components. First, on the basis of a comparative study of multiple translations of the same document into a single target language, a pragmatics-based computational model is to be developed in which reasoning about the beliefs of the participants in the translation task and about the content of a text are central. Second, existing Natural Language Processing technologies are to be appraised as potential components of a computational platform that supports investigations into the effects of pragmatics on translation. Third, the platform is to be assembled and prototype translation systems implemented which conform to the pragmatics-based computational model of translation. Finally, a novel evaluation methodology is to be developed and evaluations of the systems carried out.

Introduction

As early as 1959, Bar-Hillel (1960) pointed out that machine translation requires knowledge of the world and inferencing on the basis of that knowledge, if high-quality translation is to be achieved. Currently, only a few translation systems attempt to exploit such knowledge or knowledge-based inferencing in producing their translations. Beyond this, Farwell and Helmreich (1995), among others, have pointed out that high-quality MT further requires the ability to model the beliefs of the various participants in the translation task. To date, however, no translation system exploits beliefs ascription or participant modeling in determining its translations.

The objectives of the program of research we are proposing here are to establish a framework and methodology for investigating the pragmatic aspects of the translation process and to assemble a computational platform for carrying out systematic experiments on the pragmatics of translation. If successful, it will serve as a foundation for a research program whose ultimate goal is fully automatic, high-quality MT and whose intermediate results should improve current knowledge-based MT systems as well as other NLP applications. The core observation is that to improve MT quality significantly, the attention of the research community must shift from manipulating linguistic forms to correlating those forms with propositional content and to reasoning about that propositional content in the context of the goals and strategies of participants in a language event.

Vision of Future

We assume that the translation process consists of assigning to each successive utterance of a source language communication some representation of propositional content based on the form of the expression and given the information in the context of the utterance. The resulting interpretation not only includes the semantic content of the expression uttered but it also includes the beliefs and inferences which led to establishing that particular semantic content, along with its illocutionary and perlocutionary intent.

It is this entire interpretation that is used for translation since the object of translation is to recreate, to the extent possible, the original source language activity. Thus, if certain knowledge or a particular inference is needed to identify a participant's intent in the source language interaction, then it should also be relied on in the target language interaction if possible. At the same time, any beliefs or inferences based on assumed shared knowledge of the source language speech group must be replaced by corresponding knowledge of the target language speech group. Should no equivalent corresponding knowledge exist, communicating the explicit or implicit propositional content derived from that knowledge requires an alternative strategy. In any case, after establishing a correspondence between the central propositional content and target language expressions, the result is manifested through text or speech.

As an example, consider the translations of the Spanish expressions *del tercer piso* and *el segundo piso* in the text fragment below which is from a news article about the Moscow real estate market in the early 1990's (from Farwell and Helmreich 1997a):

> *...los 300 metros cuadrados d**el** **tercer piso***
> ...the 300 square meters on-the third floor
> *estaban disponibles pero fueron aquilados ...,*
> were available but they-were rented-out ...,
> *sólo queda **el segundo piso**...*
> only remains the second floor...

In one case, a translator rendered these expressions as *the third floor* and *the second floor*, respectively, while, in another, a different translator rendered them as *the fourth floor* and *the third floor*. Both translations are

appropriate and potentially accurate. They arise from a specific difference in beliefs between the translators about the floor-naming convention in use by the addressees of the original text, on the one hand, and by the addressees of the translation, on the other.

Assuming that a computational platform has been assembled and a pragmatics-based translation engine implemented, it is possible to envision how the Spanish text might be processed. The morphological, syntactic, and semantic components identify the semantic content of the two phrases (*el segundo piso*, *el tercer piso*) as something like {ιx: FLOOR(x) ∧ SECOND(x)} and {ιy: FLOOR(y) ∧ THIRD(y)}. Additional information from ontological entries for SECOND and THIRD indicates that these concepts require a starting point. There is no information in the text itself as to what that starting point is and so the system must infer one in order to establish an interpretation of the text. The discourse or background knowledge context might provide two candidates in this instance (the ground-level floor or the floor above it) corresponding to two different floor-naming conventions. This background knowledge also provides information about the sorts of people who use these conventions. At the same time, information from the utterance context (e.g., that the article was written in Spanish, that the person quoted was a Russian-named businessman in Moscow, etc.) can be used by a default inferencing engine to assess which floor-naming convention was being used by the author and, therefore, to which floor the phrase in question is actually used to refer. The beliefs and other information used by the system in reaching this conclusion are now made part of the utterance context (and, indeed, of the interpretation of the utterance itself).

Translation, then, involves formulating a target-language expression which the target language audience can use to construct an identical interpretation or, at least, one that is as similar as possible to the original interpretation. However, the beliefs about floor-naming convention used by the target language audience may result in the selection of semantically different expressions in constructing the translation.

What all this makes clear is that the beliefs of the participants in the translation process, as well as those of the people involved in the events described in the text itself, contribute to determining an interpretation of the original text. Beyond that, they contribute to determining the particular target language rendering of that interpretation.

Any system which supports pragmatics-based processing, then, needs to be able to model the background beliefs of the participants in the translation process, that is, the author and addressees of the source language activity and the addressees of the translation or target language activity. It also needs to model the context of active information for each successive utterance making up the source and target language activity. In order to do this, the system must have the capability of representing propositions about the world unambiguously and using those propositions as a basis for inferring information that is intended by some participant to be inferred. Finally, the target system must have individual language components that relate expressions to propositional content and vice versa.

Motivation

Clearly form-based approaches cannot succeed in producing fully automatic, high-quality translation. That is to say, any method of machine translation based solely on linguistic form and cross-lingual correspondences of form, which does not take into account the informational context and speaker's goals, can achieve only partial success. At the time Bar-Hillel made his central observation, he also suggested research into practical MT should focus on automated techniques for circumventing the translation process or for artificially limiting the class of texts to be processed. Thus, much research into MT has usually aimed at developing a "bag of tricks" intended to achieve the same input-output performance without concern for what the "natural" process was. This led to the development of systems such as METEO (Chandioux 1976) that focus on translating very restricted text types or sublanguages (Kittredge & Lehrberger 1982), systems such as the Systran application at Xerox (Ruffino 1982) and the Kant application at Caterpillar (Nyberg & Mitamura 1995) that translate texts whose creation has been strictly controlled for formal properties, i.e., controlled languages, and systems that produce low-quality, albeit potentially useful, texts using a variety of form-based statistical approaches such as Candide (Brown *et al*. 1993) or Gazelle/ReWrite (Koehn & Knight 2000, Knight & Al-Onaizan 1998) and example-based approaches such as (Collins *et al*. 1996; Furuse & Iida 1992). It also led to systems that were intended to be used in conjunction with people, principally translators, focussing on improving an organization's translation throughput. Most of these were Machine-Assisted Human Translation (MAHT) systems such as Lernout & Hauspie's T1 Professional (Schwall & Thurmair 1997) and TransSearch (Macklovitch *et al*. 2000) although some were Human-Assisted Machine Translation systems such as METAL (Thurmair 1990) and Pangloss (Nirenburg *et al*. 1995).

All of this research addresses the development of "practical" MT systems, that is, systems that produce incrementally-improved translations in the short term. Research into long term solutions which attempt to model the process of translation as it appears to be done by humans, however, have been limited. There was a certain amount of work in the area in the 1970's in AI (e.g., Wilks 1973, Carbonell *et al*. 1981, Nirenburg *et al*. 1985) and more recently within Computational Linguistics (e.g., Nirenburg 1989, Hobbs & Kameyama 1990, Dorr *et al*. 1995, Al-Onaizan *et al*. 2000, Farwell & Helmreich 1999).

Beginning around 1990, Farwell and Helmreich began to work jointly on a series of papers directed at developing a pragmatics-based model of translation (for summary, see Farwell & Helmreich 1999). Farwell and Helmreich (1995) argued for the necessity of translation based on a ramified source language analysis. Such an analysis not only includes the semantic content of the utterance (which, coopting Austin (1975), may be referred to as the locutionary intent), but also inferred information such as implied informational content, the motivation of the speaker (illocutionary intent) and the desired effect the speaker wished to have

(perlocutionary intent). Any or all of this additional information might serve as the primary justification for choosing a particular target language translation.

A comparative analysis of multiple (2) English translations of Spanish news articles led to the recognition that most variations in human translation arise from different beliefs on the part of the translators either about what the author's worldview is (Helmreich & Farwell 1998) or about what the worldview of the addressees of the translation is (Farwell & Helmreich 1997a). A model of translation was developed which recognizes a two-fold notion of context. The first of these is the discourse context, i.e., the translator's beliefs about the world and about the source and target language cultural conventions. The second is the utterance context, i.e., the translator's beliefs about the author, the addressee and the individuals and events mentioned thus far in the specific text being translated. The former exists prior to and forms the background knowledge for any discourse while the latter is constructed as the discourse is processed utterance by utterance. These contexts supply the beliefs that are used in creating the various beliefs-environments of the participants in the translation process. The environments, in turn, are used to infer the ramified analysis (interpretation) of a source language input as well as to formulate the target language rendering of that interpretation.

Finally, Helmreich and Farwell (2000) shows how an interpretation can be readily represented using the system of Text Meaning Representation (TMR-Mahesh & Nirenburg 1996). In addition, the TMR components can be used to direct the inferencing process, thus limiting the type and amount of inferencing to be done.

Path to follow

A program of research into pragmatics-based Machine Translation (MT) includes four broad components:

- developing a pragmatics-based processing model of translation,
- assembling the computational infrastructure for implementing such a model of translation,
- implementing one or more experimental pragmatics-based MT systems,
- testing and evaluating the systems and assessing the adequacy of the processing model.

We suggest that the initial component consist of a comparative analysis of multiple translations into a particular language of various multilingual corpora of texts. Such a data set would be akin to that assembled by White *et al*. (1994) for MT evaluation. The translations are compared to identify and categorize any meaning-bearing differences. These differences will be accounted for in terms of the beliefs of the translators about the world, about the beliefs of the various participants in the translation process or about the content of the text, making explicit the reasoning which led to them. The results will provide a basis for the development of a pragmatics-based computational model describing the translation process in which reasoning about the beliefs of the participants in the translation task and about the content of the text form a central component.

Second, and at the same time, existing Natural Language Processing technologies will be appraised as potential components of a computational platform that will support investigations of the effects of pragmatics on translation. We would expect such a system to be capable of (a) producing meaning representations from linguistic form using standard NLP components, (b) non-monotonic (defeasible) reasoning over incomplete data using an inferencing engine akin to ATT-Meta (Barnden *et al.* 1996) and (c) representing and manipulating embedded beliefs-contexts of various degrees of complexity using an ascription system such as ViewGen (Ballim & Wilks 1990). In addition, large-scale static resources would be required including an ontology, e.g., ONTOS (Mahesh & Nirenburg 1995) or Cycorp's Cyc (Lenat 1995), a factual knowledge database, e.g., the CRL's Fact Database (Sheremetyeva, *et al*. 1998), as well as traditional NLP knowledge sources such as lexicons and grammars. Various components will then be selected and a platform assembled.

Third, prototype translation systems from the different source languages used for the comparative study into the particular target language will be implemented which conform to the pragmatics-based computational model of translation. This step also includes the acquisition of underlying domain knowledge and linguistic conventions.

Finally, while the traditional method of rating the MT output on the basis of criteria such as fluency, adequacy and comprehension would be applied, a novel evaluation methodology must be developed which relies on comparing the actual beliefs, facts and inferences used by translators in producing their translations to those of the system. This would be extended by looking at outputs produced by a system under differing configurations beliefs-context.

Bibliographical References

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., Yamada, K. 2000. Translating with Scarce Resources. In *Proceedings of the American Association for Artificial Intelligence conference* (AAAI'00).

Austin, J. L. 1975. *How to Do Things with Words.* Cambridge, MA: Harvard University Press.

Ballim, A., and Y. Wilks. 1990. *Artificial Believers: The Ascription of Beliefs.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Barnden, J., S. Helmreich, E. Iverson, and G. Stein. 1996. Artificial intelligence and metaphors of mind: within-vehicle reasoning and its benefits. *Metaphor and Symbolic Activity*, 11(2): 101-123.

Bar-Hillel, Y. 1960. The present status of automatic translation of language. *Advances in Computers*, 1: 91-163.

Brown, P., S. Della-Pietra, V. Della-Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263-312.

Chandioux, J. 1976. Météo: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. *META*, 21, 127-133.

Carbonell, J., R. Cullingford, and A. Gershman. 1981. Steps toward Knowledge-based Machine Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3: 376-392.

Collins, B., P. Cunningham, and T. Veale. 1996. An Example Based Approach to Machine Translation. *Expanding MT Horizons: Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, Montreal. 125-134.

Dorr, B., J. Garman, and A. Weinberg. 1995. From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9(3): 71-100.

Farwell, D., and S. Helmreich. 1995. This is Not a Bedroom Farce: Pragmatics and Translation. *Proceedings of the 4th International Colloquium on Cognitive Science*, San Sebastian, Spain, 73-83.

Farwell, D., and S. Helmreich. 1997a. What floor is this? Beliefs and Translation. *Proceedings of the 5th International Colloquium on Cognitive Science*, San Sebastian, Spain, 95-102.

Farwell, D., and S. Helmreich. 1997b. Assassins or Murders: Translation of Politically Sensitive Material. Paper presented at the 26th Annual Meeting of Linguistics Association of the Southwest, University of California at Los Angeles, Los Angeles, CA. October, 1997.

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

Furuse, O., and H. Iida. 1992. An example-based method for transfer-driven Machine Translation. *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto. 139--150.

Helmreich, S., and D. Farwell. 1998. Translation Differences and Pragmatics-Based MT. *Machine Translation*, 13(1): 17-39.

Helmreich, S., and D. Farwell. 2000. Text Meaning Representation as a Basis for Representation of Text Interpretation. E*nvisioning Machine Translation in the Information Future: Proceedings of the 4th Conference for the Association for Machine Translation in the Americas, AMTA 2000*, Cuernavaca, Mexico, October 2000. Lecture Notes in Artificial Intelligence 1934, Subseries of Lecture Notes in Computer Science. Berlin: Springer Verlag.

Hobbs, J., and M. Kameyama. 1990. Translation by Abduction. *Proceedings of the 13th International Conference on Computational Linguistics*, 3: 155-161.

Kittredge, R., and J. Lehrberger (eds.). 1982. *Sublanguage: studies of language in restricted semantic domains*. De Gruyter: Berlin.

Knight, K., and Y. Al-Onaizan. 1998. Translation with Finite-State Devices. In D. Farwell, L. Gerber and E. Hovy (eds.), *Machine Translation and the Information Soup: Proceedings of the 3rd Conference of the Association of Machine Translation in the Americas*, 421-437. Springer-Verlag: Berlin.

Koehn, P., and Knight, K. 2000. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. In *Proceedings of the American Association for Artificial Intelligence conference* (AAAI'00).

Macklovich, E., M. Simard, and P. Langlais. 2000. TransSearch: Free Translation Memory on the World Wide Web. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens. 1201-1208.

Lenat, D. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *CACM* 38(11): 32-38.

Mahesh, K., and S. Nirenburg. 1995. A Situated Ontology for Practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence.* Montreal, Canada.

Mahesh, K., and S. Nirenburg. 1996. Meaning Representation for Knowledge Sharing in Practical Machine Translation. In *Proceedings of the Florida AI Research Symposium-96, Track on Information Interchange*, May 1996.

Nirenburg, S. 1989. Knowledge-based Machine Translation. *Machine Translation*, 4: 5-24.

Nirenburg, S., (ed.). 1995. The Pangloss Mark III Machine Translation System. Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie-Mellon University), Information Sciences Institute (University of Southern California). Issued as *CMU Technical Report CMU-CMT-95-145*.

Nirenburg, S., V. Raskin, and A. Tucker. 1985. The Structure of Interlingua in TRANSLATOR. In S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issues*, 90-113. Cambridge, UK: Cambridge University Press.

Nyberg, E., and T. Mitamura. 1995. Controlled English for Knowledge-Based Machine Translation: experience with the KANT System. *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation.*

Ruffino, J. 1982. Coping with Machine Translation. In V. Lawson (ed.), *Practical Experience of Machine Translation*, 35-58. North-Holland: Amsterdam.

Sheremetyeva S., J. Cowie, S. Nirenburg, and R. Zajac. 1998. Multilingual Onomasticon as a Multipurpose NLP Resource. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain.

Schwall, U., and G. Thurmair. 1997. From METAL to T1: Systems and Components for Machine Translation Applications. *Proceedings of the Machine Translation Summit IV*, San Diego, CA. 180-190.

Thurmair, G. 1990. Complex Lexical Transfer in METAL. *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation*, Austin, TX. 91-107.

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.

Wilks, Y. 1973. An Artificial Intelligence Approach to Machine Translation. In R. Schank & K. Colby (eds.), *Computer Models of Thought and Language*, 114-151. San Francisco: Freeman.