Machine Translation

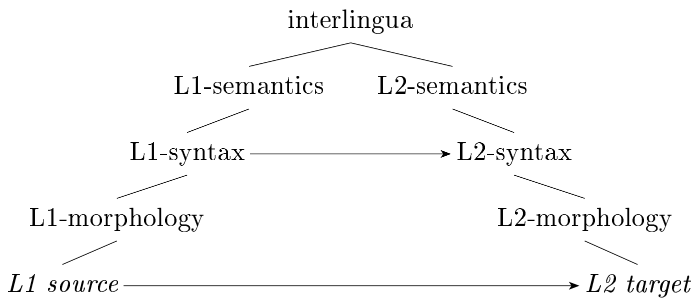
Christopher Manning

June 3, 2002

History: the success of code breaking by computers in WWII — proposed by Warren Weaver in the US — the Cold War brings copious funding — the debunking of early work by the ALPAC report and Bar-Hillel — the success of small knowledge-based systems outside the US such as METEO and GETA (Grenoble) — the Japanese 5th generation project, and the need for MT in Japan — renewed interest in the US — statistical approaches using large corpora of texts — new ideas involving human-machine cooperation — the sense in which SYSTRAN is now seen as a success — the needs of the EC, the UN, and multinational corporations.

Book-length MT Introductions include: Hutchins and Somers (1992), Slocum (1988), and Arnold et al. (1994).

How to Translate: Analysis, transfer, and interlingua (Vauquois' Triangle)



Advantages of interlingua are reducing nm to $n+m$, and the research goal of finding some universal conceptual structure. Disadvantage is the difficulty (or perhaps in principle impossibility) of doing this: every distinction in any language must be represented (which makes the ambiguity problem worse). E.g, Dutch distinguishes two words for *runway*: one for taking off, and one for landing (*startbaan* vs. *landingsbaan*). This leads to the perception that translation is by its very nature an exercise in comparative linguistics best approached by transfer.

Why is even semantic translation difficult? Word ranges differ. Cf. English *chair* with French *chaise* and *fauteuil*; or French *porte* and English *gate* and *door*. Subtle contextual factors show which translation is appropriate (cf. *We found the glove in that chair* vs. *We found the glove on that chair*).

Why is MT difficult?

Sentences have a basic overt meaning ("semantics"). In principle that should be easy to translate. But there are two problems.

1. That basic meaning may be hard to determine because of the extreme ambiguity of natural language. We saw these problems while discussing word sense disambiguation: e.g., *fencing* may be something you do with swords, or posts and wire. In most languages the two would be translated differently. Or consider bilingual dictionary entries (here Spanish-English):

acusar: to accuse; to acknowledge (receipt); to show

de: of; from; for; by

mano: hand; forefoot; coat (of paint)

pasar: to pass; to cross; to surpass; to hand; to transfer; to smuggle; ...

Or translating *(to) put on, wear* into Japanese:

kiru (general clothes); **haku** (shoes); **kakeru** (glasses); **kaburu** (hats); **hameru** (gloves); **haoru** (coats); **shimeru** (scarves, ties)

Not only is there this lexical ambiguity, but there is also syntactic and semantic ambiguity (as we have discussed). The ambiguity and context-dependence of language is actually central to its functionality – otherwise language use would be incredibly long-winded and boring, and vocabularies would have to be mind-bogglingly large. The understated structure of language allows people to succinctly state the main point, and leave the listener to infer all the relevant contextual information and assumptions.

2. Structural differences been languages:

(1) Sam wa Kim sika mi-nakat-ta
Sam TOP Kim EXCEPT saw-NEG-PAST
'Sam only saw Kim'/'Sam didn't see anyone apart from Kim.'

3. The basic meaning is often greatly removed from the interpretation of the sentence in a certain textual or real world context. In translation, we want to find an idiomatic way of expressing something in the other language which has the same contextual interpretation. Finding something with the same semantics is usually worse than useless. For example, to make an arrangement to visit someone in the afternoon in Japan, one might say *Gogo, ojama shite mo yoroshii deshoo ka?*, a semantic (often called *literal*) translation of which might be *[This] afternoon, (even) if I am a hindrance/obstruction, will it be okay [good]?*. The correct way to express something with the appropriate contextual interpretation in English might be *May I come and see you this afternoon?* which has very little lexical or syntactic overlap with the original. One might well say that this is just an idiomatic expression and the translation can be stored, and that is certainly true, but this context-dependence occurs much more widely. The French sentence *Où voluez-vous que je me mette?* means literally *Where do you want me to put myself?*, but an appropriate translation might be *Where do you want me to sit/stand/sign?* or for the first two of those interpretations (only!), maybe just *Where do you want me?* What remains invariant under translation is not the meaning, but the contextual interpretation, that is, the response the text is intended to evoke in a reader. Often the meaning of the translation is greatly different.

4. Interpretation depends on contexts, and finding a good translation may involve the inclusion of all sorts of material not in the source text (or perhaps more easily, the selective deletion of material that is there). Among the simplest but often most pressing problems is choosing which determiner to use when translating from, say Japanese to English, or what translation to choose for a pronoun when translating from French to English, or when to put in singular/plural distinctions when translating from a language that lacks them. In practice, these issues are so difficult that they are addressed fairly little (in the hope that they don't matter in more "technical" domains). But even here many of the problems involve not linguistic knowledge but common-sense and world knowledge. A sign that says "Open" on the door of a shop means that it is currently open; a waving flag saying "Open" on the pavement nearby probably means that it is a new store that is now open for business. These two cases would be translated differently in some languages (e.g., German). Language is situated; meaning underdetermines interpretation.

Summary

Fully automatic high quality translation is unrealistic for the foreseeable future.

Thus there has been much exploration of human-assisted machine translation (HAMT), and machine-assisted human translation (MAHT). Translator's workbenches which particularly use *translation memories*, which record how things were translated before, have had the greatest commercial impact, and have delivered good cost savings (and human translators like them!). There are a number of places where low quality is useful, however: screening documents, the web.

Cheap translation for the web

One of the main suppliers of cheap web translation is Systran (an old line traditional MT company):

> http://www.systransoft.com/

You can find other translation systems on the web, commonly using the above technology, including:

> http://babelfish.altavista.com/translate.dyn
> http://www.freetranslation.com/ (Claims to be "the fastest" – not what I found.)
> http://www.google.com/language_tools?hl=en

It's quite fun to do a small evaluation of these translation tool (either by making use of another European language you know, or by using the trick of "double translation").

Example (from the front page of a 2000 *NYT*:

> Later in the week, President Clinton holds his first summit in Russia with Vladimir Putin.
>
> Plus tard Prsident Clinton tient le sien dans la semaine en premier sommet en Russie avec Vladimir Putin.
> more late President Clinton holds his in the week in first summit in Russia with Vladimir Putin
>
> Later President Clinton holds his in the week, in first summit in Russia with Vladimir Putin.

Statistical MT

Pioneered by IBM in the late eighties and early 1990s.

$$\boxed{\begin{array}{c}\text{Language Model}\\ P(e)\end{array}} \xrightarrow{e} \boxed{\begin{array}{c}\text{Translation Model}\\ P(f|e)\end{array}} \xrightarrow{f} \boxed{\begin{array}{c}\text{Decoder}\\ \hat{e} = \arg\max_e P(e|f)\end{array}} \xrightarrow{\hat{e}}$$

The Language Model generates an English sentence e. A language model can be a trigram model of the sort we have seen. The Translation Model transmits e as the French sentence f. The decoder finds the English sentence $\hat{e}$ which is most likely to have given rise to f.

Decoder. This one does the standard kind of maximization, based on the observation that we can omit $P(f)$ from the maximization since f is fixed:

$$\hat{e} = \arg\max_e P(e|f) = \arg\max_e \frac{P(e)P(f|e)}{P(f)} = \arg\max_e P(e)P(f|e) \tag{1}$$

Translation model. Here is a simple translation model based on word alignment:

$$P(f|e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(f_j|e_{a_j}) \tag{2}$$

We use the notation of Brown et al. (1993): e is the English sentence; l is the length of e in words; f is the French sentence; m is the length of f; f_j is word j in f; a_j is the position in e that f_j is aligned with; e_{a_j} is the word in e that f_j is aligned with; $P(w_f|w_e)$ is the *translation probability*, the probability that we will see w_f in the French sentence given that we see w_e in the English sentence; and Z is a normalization constant.

Translation probabilities. The translation probabilities are estimated using the EM algorithm. One way to guess at which words correspond to each other is to compute an association measure like χ^2. But that will generate many spurious correspondences because a source word is not penalized for being associated with more than one target word.

The basic idea of the EM algorithm here is that it solves the *credit assignment* problem. If a word in the source is strongly aligned with a word in the target, then it is not available anymore to be aligned with other words in the target.

We start with an initialization of the translation probabilities $P(w_f|w_e)$ based on uniformity given sentence-level alignment. In the E step, we compute the expected number of times we will find w_f in the French sentence given that we have w_e in the English sentence.

$$z_{w_f,w_e} = \sum_{(e,f) \text{ s.t. } w_e \in e, w_f \in f} P(w_f|w_e)$$

where the summation ranges over all pairs of aligned sentences such that the English sentence contains w_e and the French sentence contains w_f. (We have simplified slightly here since we ignore cases where words occur more than once in a sentence.)

The M step reestimates the translation probabilities from these expectations:

$$P(w_f|w_e) = \frac{z_{w_f,w_e}}{\sum_v z_{w_f,v}}$$

where the summation ranges over all English words v.

What we have described is a very simple version of the algorithms described by Brown et al. (1990) and Brown et al. (1993). See also Berger et al. (1994). Fairly recent work takes some account of structure – see the refs at the start of the further reading of chapter 13 (pp. 492–493). For very recent work on statistical MT, one of the best sources is Kevin Knight's webpage:

http://www.isi.edu/natural-language/people/knight.html

Bibliography

Arnold, D., L. Balkan, R. L. Humphreys, S. Meijer, and L. Sadler. 1994. *Machine Translation: An Introductory Guide.* Blackwell.

Berger, A. L., P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, and L. Ureš. 1994. The Candide system for machine translation. In *Proceedings of the 1994 Human Language Technology Workshop*, 157–162. DARPA.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2):79–85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Hutchins, W. J., and H. Somers. 1992. *Introduction to Machine Translation.* Academic Press.

Slocum, J. 1988. *Machine Translation Systems.* Cambridge University Press.