Generation Issues in Machine Translation

Gregor Thurmair
Linguatec
Gottfried-Keller-Str. 12
Munich
g.thurmair@linguatec.net

Abstract

Generation in MT must be based on meta-structure information on the author's communicative intentions, which must be decoded from the source language text.

1 Generation in Machine Translation

The history of generation in MT can be seen as a decline in understanding of the generation task.

In the *rule-based paradigm*, generation in MT, after having mastered constituent-order problems and inflection, turned its focus to phenomena which are not easily shared between languages, or are context-sensitive / supra-sentential.

Examples of the first type are translations of tense, aspect or definiteness, like:

- How to determine aspect of Russian verbs in a German-Russian system (aspect is not overtly marked in German) (Klimonov 1991)
- How to decide on the definiteness of noun phrase determiners in a Russian-German system (Russian does not have determiners)

Examples of the second type was research on topic and focus, and how to recognize and translate such phenomena (Steiner / Winter-Thielen 1988).

This research was not really continued later.

In the area of *unification-based approaches*, the idea was to use the constraints of the lexical elements of a sentence, and apply them to generation just like to analysis. The paradigm was called 'shake-and-bake translation', assuming a bag of words which would organize itself into a well-formed target structure once all constraints of the lexical elements are satisfied. This approach has been re-vitalized recently in the METIS project (Carl et al., 2005).

The approach proved to work only for selected lexical elements but not for real-life sentences. Also, the words themselves do not determine e.g. whether they should form a declarative or an interrogative sentence, which shows that meta-information for generation, beyond pure lexical information, is required.

In the era of *example-based MT*, generation was neglected in system design. Effort focused mainly on identifying examples in the source text, esp. bilingual term and phrase detection. How to compose such elements in a meaningful way in the target language was not researched. A good example is Trujillo (1999): In the chapter on example-based approaches, a section on generation is simply non-existent.

In the current era of *statistical MT*, generation uses language models, mainly bigram or trigram models (Knight / Koehn 2003). The only source of knowledge used is co-occurrence probability, usually derived from training corpora.

These corpora may not match the domain, nor the text structure, nor communicative intentions of the text to be generated. The results are not convincing at all, esp. if morphologically richer languages, like German, Russian or Greek need to be generated, and lead to the conclusion that more intelligent generation approaches should be used.

2 Interface representation

It is a commonplace in linguistics that the meaning of a sentence is determined not just by its words (word semantics) but also by its structure (sentential semantics). Both components encode a communicative intention of the text author. In

translation, it is this communicative intention which must be carried over into the target language.

Accordingly, (target) language generation must be seen as providing not just the propositional content, but in addition meta-information on communicative and textual aspects. Such information would be available in the form of feature structures, in MT sometimes called "transfer interstructures" (Alonso 1990). From there, different utterances should be able to be created. (Test of a generation component, ideally, uses a test frame which allows to systematically manipulate this transfer interstructure, by changing feature values (singular>plural, declarative> interrogative etc.), and evaluating the generation output.).

The question is which meta-information is required for good generation output, and where this information can be found.

3 Generation meta-information

3.1 Which information is needed

To generate a sentence / text, several layers of information must be provided:

- the words which should participate in the generation, and word-related information. In MT, this information comes from dictionaries.
- grouping the words into phrases, and assigning properties to those phrases: definiteness for noun phrases, tense / aspect for verb phrases
- relating these phrases into propositional structures, esp. syntactic functions (subject – object). Without this, proper generation of case marking in German or Greek is impossible.
- Sentential properties (sentence mood; focus)
- Pragmatics (text coherence, elliptic structures) and speakers' attitudes towards the propositional content (irony, unbelief etc.).

The less information is available the less accurate a translation will be: If only words are available then there is no help but to order them somehow. If no information on definiteness on a noun phrase is available, some default needs to be generated, which tends to mess up understanding significantly. If syntactic functions are available then at least case marking of subject and object etc. can be properly set.

Different MT systems can be ranked according to the information they have available for generation. Of course the human-like perfect generation does not exist, but systems can improve.

3.2 Where the information can be found

Language generation usually is divided into two parts, a strategic / planning part, determining *what* should be said, and a tactical part, determining *how* it should be said. All information just mentioned must be provided by the strategic phase before the tactical generation can start.

Ritzke (2000) has shown that the tactical part of an MT system (constituent ordering, word inflection etc.) can be taken over into other domains, like building abstracts from results of information extraction.. The issue is planning.

The specific situation of MT is that this information, resulting from the planning phase, must be extracted from the source text, in the source analysis phase. The analysis phase needs to decode all information for planning, determine syntactic functions, determine definiteness, tense and aspect, topic and focus, etc. In the best case, analysis can fill all the parameters in the transfer interstructure required by good generation. The better the analysis is the better the overall system performance will be.

It is in this respect that analysis can be seen to be the strategic component of language generation in machine translation.

References

Alonso, J. A., 1990: "Transfer interstructure: designing an 'interlingua' for transfer-based MT systems." In Proc. 3rd TMI Conf., Austin, Tx.

Carl, M., Schmidt, P., Schütz, J., 2005: Reversible template-based Shake and Bake Generation. Proc. MT Summit 10, Phuket.

Klimonow, G., 1991: Zur Wahl des Verbalaspekts bei der Übersetzung ins Russische. Sprache und Datenverarbeitung 15,1

Knight, K., Koehn, Ph., 2003: Introduction to Statistical Machine Translation. Tut. MTSummit, New Orleans

Ritzke, J., 2000. SEN-DNL-Gen: Dynamic Natural Language generation within the SENSUS system environment. Sensus Report.

Steiner, E., Winter-Thielen, J., 1988: On the semantics of focus phenomena in EUROTRA. Proc. COLING

Trujillo, A., 1999: Translation Engines: Techniques for Machine Translation. Springer