TT2 - TransType2 (IST- IST-2001-32091)

Antonio S. Valderrábanos*          Luis Iraola Moreno*
antonio.valderrabanos@madrid.sema.slb.com luis.iraola@madrid.sema.slb.com

Jose Esteban Lauzán*
jfernando.esteban@madrid.sema.slb.com

*SchlumbergerSema sae
Calle Albarracín, 25
28037 Madrid - Spain

1. Position with respect to existing paradigms

Within the area of Translation Automation, it is customary to distinguish between:

1.  Fully Automatic Machine Translation (FAMT) systems, and

Traditional MT systems (e.g. Systrans) belong to the FAMT group, and they vary with regard to the complexity of the input text they can understand (from free text to highly controlled languages) and to the quality of the output text they can produce (from “gist” translations to near-to-zero post-editing effort).

MAHT systems do not attempt to produce target texts without the intervention of the human translator but to help him/her in achieving higher productivity rates while increasing the final quality. The first goal is commonly achieved by building a translation memory, a (usually large) base of source-to-target text fragments that the system tries to employ when newly (though similar) source text is presented to the system. Consistency, at the sentence level, is ensured by reusing previous (approved) translations; at the terminological level, MAHT tools help translators to build and share terminological databases. Terminology management tools are included almost in every application that attempts to automate the translation process, particularly in MAHT applications but also in MT systems.

The TT2 system could be considered as a CAT tool, in the broad sense of the term, since it is conceived as a translation aid that attempts to increase productivity and quality in translation. However, it departs from current CAT tools in the technology employed for achieving those goals. Commercial CAT tools employ previous translations (translation memories) for suggesting new translations and these translations are exploited via string-matching procedures (either exact or fuzzy). In contrast, TT2 exploits previous translations using statistical MT techniques; as a result TT2 is not tied to exact or fuzzy matches at the sentence level. TT2 can be described as an interactive machine translation (IMT) system.

2. Objectives

The objective of TransType2 (TT2) is building a software aid for human translators. The project is planned as a continuation of the work of TransType, a project successfully carried out by one of the partners of the current consortium, the RALI Laboratory of the University of Montreal.

The originality of the TransType approach consists on helping the translator while he/she types the translation, so taking into account the source text and the part of the target text already typed in order to propose sensible completions. Typically, more than one completion for the text already typed is proposed by the system, which orders its
proposals from the most likely completions to
the most improbable ones. The user selects
among the proposed completions or, if none
of them is judged correct, continues typing
the desired translation. The tool aims at
reducing the amount of typing needed for
completing the translation by proposing
translation pieces that after being validated by
the translator are entered into the target text.

The new project maintains the original
approach while enhancing the capabilities of
the initial prototype in three directions:

• More translation pairs are added:
  • French to English and vice versa
  • German to English and vice versa
  • Spanish to English and vice versa
• Longer and more accurate translation
  pieces are proposed to the translator
• A new interaction modality, speech,
  is introduced and extensive user
  testing is planned in order to increase
  the usability of the system.

In order to present to the translator correct
completions of the target text being typed, the
TT2 system employs a language model for
the required source-target translation pair.
This language model is the result of the
statistical analysis of large parallel corpora of
source-target texts in which translation
 correspondences are learned.

The technology that underlies TransType is
basically composed of two probabilistic
models: first, an interpolated trigram model
of the target language; and second, a
translation model that essentially corresponds
to IBM Model-2. The two models are
combined linearly in the current prototype,
using a fixed weighting coefficient. To
predict longer sentence fragments, however,
TT2 will incorporate more sophisticated
translation models that extend beyond word-
for-word substitution.

TT2 also builds on a previous project,
EuTrans, where partners of TT2 where
involved, RWTH Aachen of the University of
Technology and the Instituto Tecnológico de
Informática, Universidad Politécnica de
Valencia. Although focussed on fully automatic translation, the experiences from
EuTrans in the field of stochastic and finite-
state methods for machine translation have
provided an ideal starting point for TT2. In
terms of translation modeling, EuTrans
showed that Models 3 to 5 and extensions
thereof, like Alignment Templates, have
resulted in better alignment and translation
quality. This is also the case for the
incorporation of morpho-syntactic analysis.

In TT2, EuTrans is being improved in
different directions. The domain to which the
technology will be applied is going to be
significantly less restricted, compared to that
of EuTrans tasks (like the "Traveler Task").
This change will have two major
consequences:

• The size of the corpora and
  vocabulary that TT2 handles will be
  much larger. The corpus is expected
to be 1.000.000 words per language
  and the vocabulary will be of about
  20.000 word forms.
• The complexity of the linguistic
  structures considered is higher.

Besides, data driven MT technology will be
applied in an interactive translation
environment. Instead of providing one single
full translation per sentence, the MT engine
will interactively modify its candidate
translation for a given sentence, according to
the target text that the translator is
composing. We could say that the MT engine
would be maintaining a "continuous
dialogue" with the translator. Facilitating this
interaction between the translator and a MT
system is one of the most innovative and
promising aspects of TT2.

1 Note of the Authors: TT2 started only 5
months ago and the results achieved so far
(about issues like Market Analysis or User
Specifications) are not mature enough to be
presented at this point (July 2022).