

Discovering MT strategies
beyond word-for-word translation:
a laboratory assignment

Juan Antonio Pérez-Ortiz and Mikel L.
Forcada

Departament de Llenguatges i Sistemes
Informàtics

Universitat d'Alacant, E-03071 Alacant
(Spain)

Index

- ⇒ A common misconception
- ⇒ A note on “word-for-word”
- ⇒ “Model zero”
- ⇒ “Model zero” from commercial systems
- ⇒ A laboratory assignment
- ⇒ Concluding remarks

A common misconception

“*Machine translation systems translate
word for word. This is why they don't
work.*”

- ⇒ This misconception needs to be
addressed in the classroom
- ⇒ But may actually be a good starting
point!
- ⇒ However, let's clarify what “word for
word” means

A note on “word-for-word translation”

- ⇒ A confusing term (even in our paper!)
- ⇒ For some, it refers to *direct* or *transformer*
architectures which actually do a bit more
than word-for-word
- ⇒ For others, it simply refers to a brute-force
left-to-right word-substitution strategy
- ⇒ Let's focus on the second case and use
the name “model zero” for it.

The simplest word-for-word strategy: “model zero” /1

A “model zero” MT system:

- ↻ Reads the SL text word by word, from left to right
- ↻ Writes the TL text by picking a constant equivalent for each SL word and placing words in SL order.

This is a very nice reference model.

Even the cheapest MT systems do more.

F i f di

The simplest word-for-word strategy: “model zero” /2

“Model zero” is useful:

- ↻ Easy to understand
- ↻ Easy to criticize
- ↻ Nice to build upon

Expected failures of “model zero”

With “model zero” one may get

- ↻ Wrong agreement (number, gender)
- ↻ Wrong word order
- ↻ Wrong translations for homographs showing different parts of speech
- ↻ Wrong translations for multiword units (idioms, collocations, etc.)

But even the cheapest systems do better!

Using commercial systems in the laboratory

- ↻ Balkan et al.'s (1997) seminal report: using commercial systems in the classroom is beneficial
- ↻ Many systems are available at zero cost (as internet services).

Getting “model zero” behavior from commercial systems

Real, commercial systems go beyond “model zero”. So how?

One can simulate “model zero” in the laboratory by using *single-word sentences*.

(or, to be really sure, *single-word paragraphs*).

This forces MT systems to deliver a constant equivalent for each word.

A laboratory assignment/1

Assignment:

- ⇒ Give a set of (selected) sentences to students
- ⇒ Ask them to translate them with different MT systems, both as whole sentences and forcing “model zero”.
- ⇒ Ask them to compare the results and to discuss the differences in terms of a MT strategy.

A laboratory assignment/2

Commercial programs we have used (English to Spanish):

- ⇒ Globalink's Power Translator (old versions are cheap)
- ⇒ Transparent Technologies' Transcend RT (www.freetranslation.com)
- ⇒ Softissimo's Reverso (www.reverso.net)

A laboratory assignment/3

Questions to guide the discussion:

- ⇒ “Why are “model zero” and sentence translations different?”
- ⇒ “Which rules uses the system to choose equivalents for words?”
- ⇒ “Apart from choosing words, what other operations does the program perform?”
- ⇒ “Could you dare advancing an explanation of the strategy used?”

A laboratory assignment/4

Example 1

Source: My tailor is rich

Zero: Mi *adapte es rico

Sentence: Mi sastre es rico

Homograph *tailor* taken as a verb (*adaptar*) in isolation.

Strategy: Context words used by MT system to make the correct choice.

A laboratory assignment/5

Example 2

Source: Artificial intelligence systems can think

Zero: Artificial / la inteligencia / los sistemas / poder / piense

Sentence: Los sistemas de inteligencia artificial pueden pensar

Word order, addition of articles, number agreement of verb, infinitive after *can*

A laboratory assignment/6

Example 2 (cont.)

Strategies:

- ⇒ Word order: some sequences of words are reordered (Forcada, MT 2000)
- ⇒ Addition of articles to nouns to get correct noun phrases in Spanish
- ⇒ Number of verb taken from preceding noun

A laboratory assignment/7

Example 2 (cont.)

Strategies (cont.):

- ⇒ Infinitive chosen after modal verb “pueden pensar”

A laboratory assignment/8

Example 3

Source: Machine translation systems cannot translate complex texts

Zero: Elabore / la traducción / programa / no poder / traduzca / el complejo / los textos

Sentence: Los programas de traducción automática no pueden traducir textos complejos

A laboratory assignment/9

Example 3 (cont)

New phenomena:

- ↻ Multiword unit “machine translation” = “traducción automática”, not “traducción de máquina”
- ↻ Choice of adjective for “complex”
- ↻ Agreement in “textos complejos”

A laboratory assignment/10

Example 3 (cont.)

Strategies:

- ↻ multiword units in dictionaries
- ↻ Adj.-noun preferred to noun-noun in English?
- ↻ Noun-adjective agreement
[More examples in paper]

A laboratory assignment/11

Hypotheses about strategies may be confirmed and refined

- ↻ By substituting source words by words in the same category (“Your tailor is rich”).
- ↻ By changing SL words from singular to plural or viceversa to check agreement
- ↻ Etc.

A laboratory assignment /12

Conclusion: MT systems do more than “model zero” word-for word translation:

- ⇒ Try to solve homographs
- ⇒ Try to get correct agreement
- ⇒ Reorder words for correct TL syntax
- ⇒ Detect and translate multiword units as a whole

Concluding remarks

- ⇒ A laboratory assignment may help students abandon their misconception that MT programs do little more than substituting words.
- ⇒ A comparison of results for words in isolation and in sentences reveals that MT programs perform many other operations.
- ⇒ Students may advance hypotheses about the strategies used.