Almost fifty years after the (first?) ALPAC report

Gábor Prószéky
MorphoLogic & Pázmány University,
Budapest, Hungary

ABSTRACT

The presentation will provide a historical flashback of Machine Translation (MT) by reviewing the significant milestones in its development and will reflect as to what the future has in store. To start with, the early developments after the Second World War II will be outlined. Next, the presentation will elaborate that the addition of morphological, syntactic and semantic knowledge did not lead to expected improvements which in turn, triggered the ALPAC report in 1966.

In the 1990s that landscape significantly changed due to the emergence of large amount of language data (corpora) which offered new opportunities for the rise and deployment of Statistical Machine Translation (SMT). SMT has been recently enhanced by the incorporation of morphological, syntactic and semantic information but the results are still not as good as expected. The presentation will review these recent developments and will reflect as to what are the options for the EU decision makers given that high quality MT is still a desideratum…

1. Some thoughts on progress in MT — almost fifty years after the (first?) ALPAC report

1. Several years after the World War II, a new scientific plan was born: translating from one language to another by the use of modern computing devices. **The idea was simple**: converting strings of language *A* into strings of language *B*, as the cryptographic activity of the war period suggested. Support of the idea of machine translation in the United States at that time was mainly motivated by the Cold War. Decision makers in the US governing bodies were quickly convinced by the new idea. There were not too many languages in their minds: first of all translation of Russian texts into English was in the focus.

    a. When the first real translation algorithms were made, a small modification of the original idea of pure string manipulation was made soon. The reason was very simple: words of Russian have inflections at their ends, thus, an unavoidable module, namely **morphological analysis**, was added to the basic “string transforming” algorithm. The results of the modified translation systems were, however, still not good enough…

b. Some years later, partly due to research results of the early generative linguists, it became clear that **syntactic structures** of human languages must play an important role in computational language processing. The analysis phase of machine translation, therefore, started to use syntactic modules to replace the simple word reordering technique. The results of the modified translation systems were, however, still not good enough…

c. Some years later, the first attempts toward computational **semantics** arose, and meaning-oriented information was added to some machine translation systems. The argument was easy to understand: high-quality machine translation cannot be made without a sort of "understanding" the string to be translated. Unfortunately, after adding so many linguistically important modules to the basic algorithm, the final results of fully automatic rule-based machine translation systems did not become significantly better.

d. The US Government had been waiting for a rather long time, but **the expected high-level results of machine translation did not come**. This situation led to the birth of the Automated Language Processing Advisory Committee, which published the opinion of its expert members in the famous **ALPAC Report** in 1966.

2. A quarter century later, in the early nineties, another new scientific plan was born: translating with the help of statistical knowledge derived from huge text corpora that were already available at that time. **The idea was simple** again: strings of language *A* can be translated into strings of language *B*, if there is enough statistical evidence for it in the corpora. Decision makers in the EU were quickly convinced by the new idea: multilingual Europe has a lot of potential language pairs and this solution would solve the problem of huge amounts of translation tasks.

    a. A step towards application of linguistically motivated modification of the basic algorithm came soon: when some languages use magnitudes more word forms than other languages, something should be done with **morphology**. To solve this problem, factored statistical translation was introduced, and made the basic algorithm a bit more better and a bit more complicated. The results of the modified translation systems were, however, still not good enough…

    b. Syntactic constructions of the languages of the EU are quite heterogeneous, thus structurally different language pairs are rather difficult to use in the statistical machine translation paradigm: some automatic tool was needed making strings of source and target languages formally more similar to each other. This led to the birth of syntactic reordering and other sophisticated **syntax-driven** algorithms added to the basic SMT paradigm. The results of the modified translation systems were, however, still not good enough…

    c. Nowadays, there are several attempts that use some sort of **semantic information** combined with the basic statistical machine translation algorithm. Unfortunately, after adding so many linguistically important modules to the basic

algorithm, the final results of statistical machine translation systems did not become significantly better.

d. The EU decision makers have been waiting for a rather long time, but **the expected high-level results of machine translation do not come**. The question to be answered is the following: How long do the EU decision makers have to wait yet for the expected high-level results of (statistical) machine translation? And if they don't come very soon, **what's the next step**?