Some Lessons Learnt by a New Comer

Low Hwee-Boon
Institute of Systems Science
National University of Singapore

The author is relatively new to the field of MT compare with all the other speakers in the panel. Through out last few years of literature readings, talking and seeking guidance from the pioneers, and managing an industrial oriented MT R&D project, the author has accumulated some observations, and learned some lessons which forms the following viewpoints. The author would like to share some of these viewpoints and seek more comments and feedback for further accumulation and dissemination to new comers of MT .

Technology Wave of MT R&D

To bet on what will be the next wave of MT technology, one can always look at the last wave of AI technology and computational linguistics research. It is not a surprise as MT actually is a large scale AI problem. Looking back at the history, MT technology in a decade was based on the AI technology of the preceding decade : MT technology in the 60s are mostly "direct" approach which basically make used of a series of pattern matching techniques. MT technology of the 70s are mostly "transfer" based approach making used of transformation of syntactical rules. In the late 80s, semantic based approach for MT was developed making use of knowledge based technology developed by AI researchers. Some attempts to leap-frog the AI technology, e.g. the CETA interlingual experiment, had fail. Will the next wave of MT technology to be developed in the 90s based on Neural Network, Case Based Reasoning, or any other AI technology emerged in the 80s ?

Application Wave of MT Technology

Likewise, the application of MT technology seems to follow a decade after the MT technology. In fact the commercialization of MT is a good filter for a wide varieties of MT technology developed. Only those MT technology which can be scaled up, either technically in handling large number of dictionaries and linguistic rules, as well as cost effective enough in the development and maintenance of Lingware have a chance of survival. Earlier commercialisation of MT, e.g. SYSTRAN, are mostly "direct" approach based MT system. In the early 80s, several new MT entrepreneurs or start-ups have been deployed for servicing large scale English/French or English/Spanish translation. None of them used interlingual approach. The MT applications of the 90s are emerging, attempting at Multilingual capabilities with semantic and knowledge based approach.

Naive View of MT

To a new comer in the field of MT, one will most likely come across a triangular figure illustrating generations of MT technology. The base line is the first generation of "direct" approach, the middle line is the second generation of transfer approach and the tip at the top of the triangle represent the "ultimate" third generation solution : "interlingual" approach.

This illustration is too simplistic in describing generations of MT technology. It is also misleading

as it implies that the interlingual approach is the latest and "last" generation in which there will be no other newer approaches to MT after this generation. On the other hand, it has an interesting feature that reflects the reality : the effort in developing a MT system based on a certain approach is proportional to the distance travels on the diagram. Hence "direct" approach takes less effort than "transfer" approach and "interlingual" approach takes the most effort.

Realistic View of MT

A more realistic illustration of MT, in my opinion, is a Hyperbola shape, where by theoretically, the two ends never meet each other. This implies that pure interlingual will be always a dream. We can only try our best and through the accumulation of knowledge and technology, reduce the "distance" between any two languages.

Few characteristics are illustrated by the hyperbolic diagram. Firstly, there exists a series of transfer approach at different language hierarchy. The typical MT technology in the 70s are at the syntactic level and that of the 80s are position at the semantic level. At ISS, we have developed a practical MT system at the deep syntactic level.

Secondly, not only the effort is proportional to the distance travel, the further up in the language hierarchy level, the more stepper it takes and hence much more effort is required to reach higher linguistic strata. The explanation for this phenomena is quite obvious. The lower level can be restricted to the domain of language. At the semantic level, one requires knowledge of the objective world, and at a higher level, one has to deal with knowledge of the spiritual world, such as feelings, emotions.

An Alternative View for the Future

To a new comer, one may ask the following question : why would one choose to take more effort in transferring at a higher language hierarchy ? Many MT guru will provide the following reasons:

> To resolve natural language ambiguities in order to achieve higher accuracy of translation.

> To increase the shared components (analysis and synthesis) and reduce the non-shared components (transfer) between pair of languages.

For some language pairs which are linguistically quite far apart, higher level of language hierarchy is a necessity to achieve a good degree of translation. This also explains why in Japan, though there were as much MT R&D effort as the rest of the world, but widely used commercial MT system are relatively fewer than the Romanized language pairs.

On the other hand, most of the practitioners are aware that there is a penalty to pay when going up the language hierarchy in the stages of analysis : the potential of introducing errors. A better strategy is "transfer at the right level": do analysis to a level of just enough information for a successful transfer. Of course, the right level of analysis depends on individual sentences, and it varies substantially between language pairs.

A diagram to reflect this philosophy is a *Ladder*. It has the following characteristics :

Similar effort at each level of transfer

    If it takes too much effort at a lower level for transfer, then one should escalate to a higher level of analysis.

Application of Machine Learning technology

    Be it Neural Network or Case Based Reasoning, the AI technology emerged in the last decade have a common trait - learning capability. Learning capability is essential in narrowing the gap of the ladder.

Future of MT Technology

For MT system to be successful by adopting the Ladder philosophy, one has to develop MT technology in the following aspects :

    Able to determine what is the right level to transform the resultant of source language analysis to that of target language.

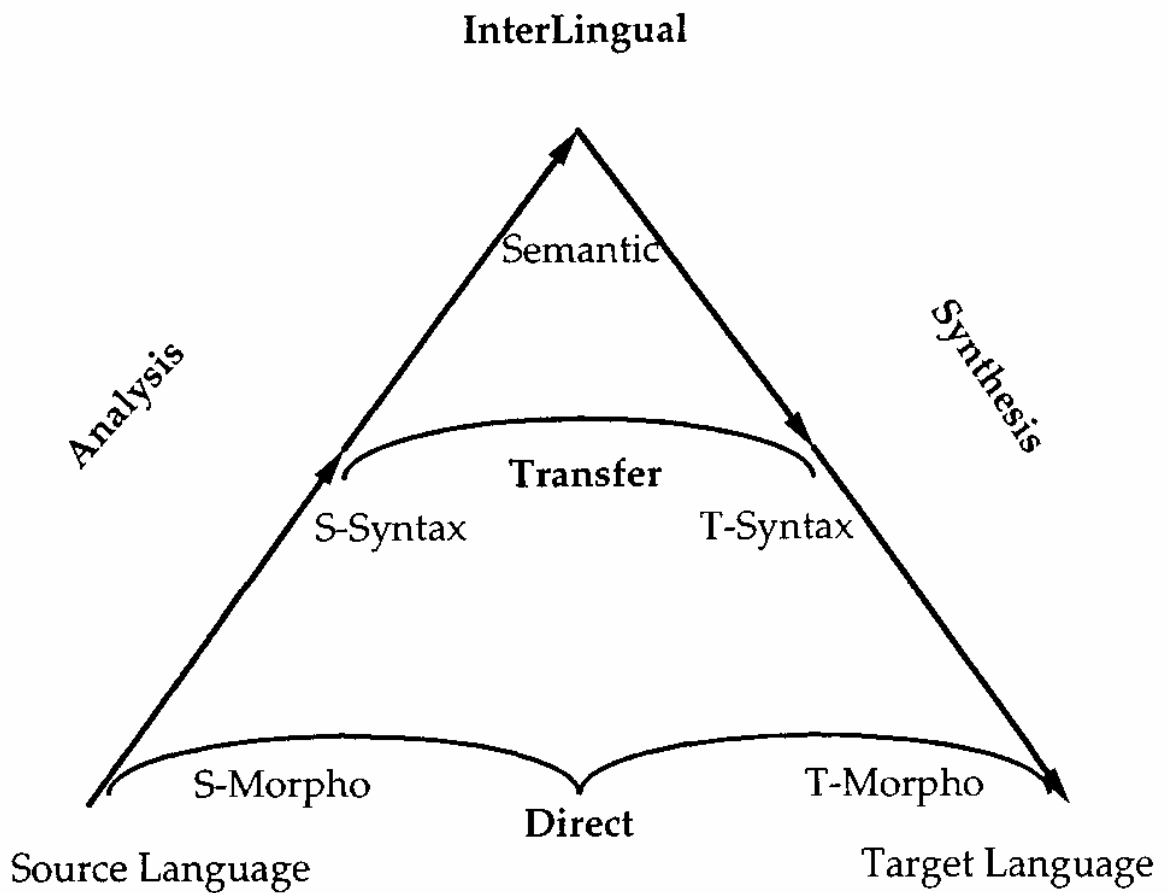   Able to minimize the effort in developing each of the transfer component.

The author has the opinion that previous transformation techniques, such as pattern matching for the direct approach, rule driven transformation for the transfer approach and even unification based interlingual approach, do not have the above two characteristics. The author believes that Neural Network paradigm could be a good substitute at lower levels and Case Based Reasoning could be appropriate for transferring at higher hierarchies.

Wave of MT Technology

AI Technology   **MT Technology**   **MT Applications**
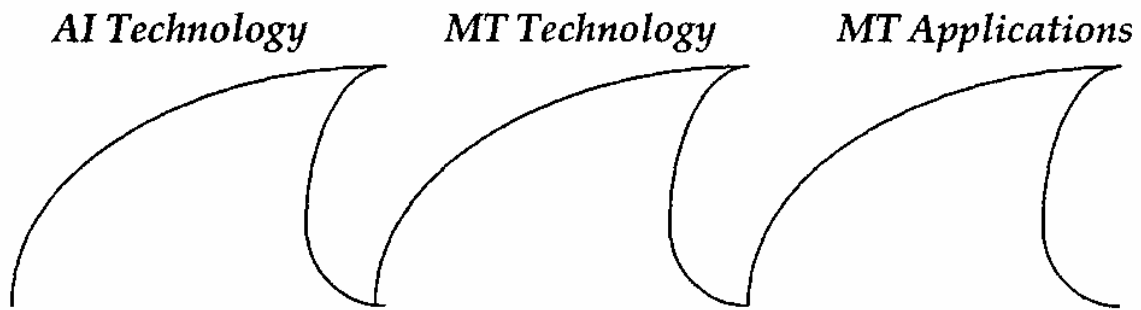
50s : Pattern Matching
60s : Parsing   Direct - Symbolic Pattern
70s : Expert System   Transfer of Syntatical Rules SYSTRAN
80s : Neural Network   Interlingual with Semantic GETA
90s : Machine Learning? NN or case based MT ?   CMU
                                                    ???

Naive View of MT

InterLingual

Semantic

Analysis

Synthesis

Transfer

S-Syntax   T-Syntax

S-Morpho   T-Morpho

Direct

Source Language   Target Language

A Simplistic View :

3 Generations of MT technology
    1st :- Direct
    2nd :- Transfer
    3rd :- Interlingual

Realistics View of MT

Analysis

Synthesis

? ? ?
Transfer

Semantic
Transfer

Deep Syntatic
Transfer

Surface Syntatic
Transfer

S-Syntax

T-Syntax

S-Morpho

Direct

T-Morpho

Source Language

Target Language
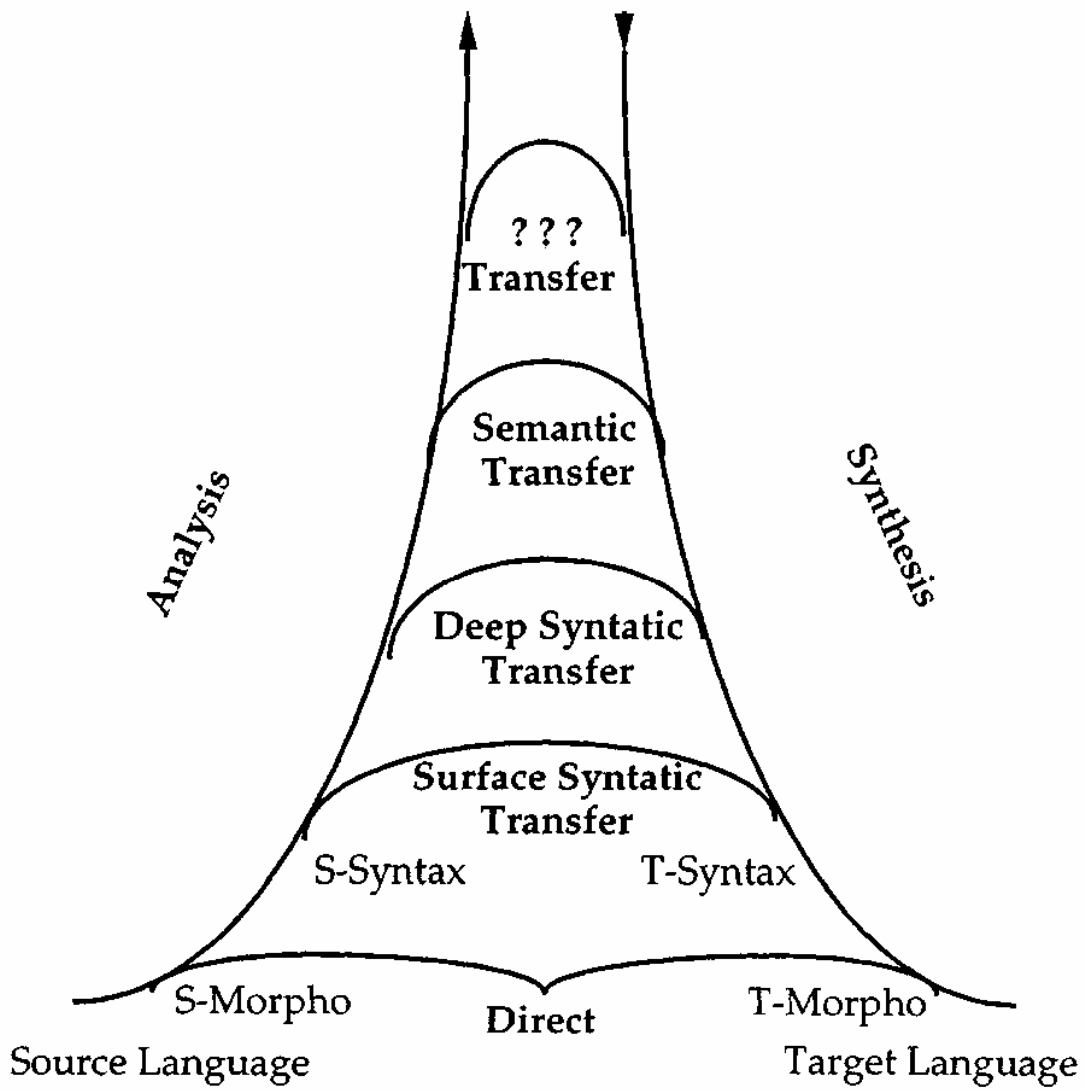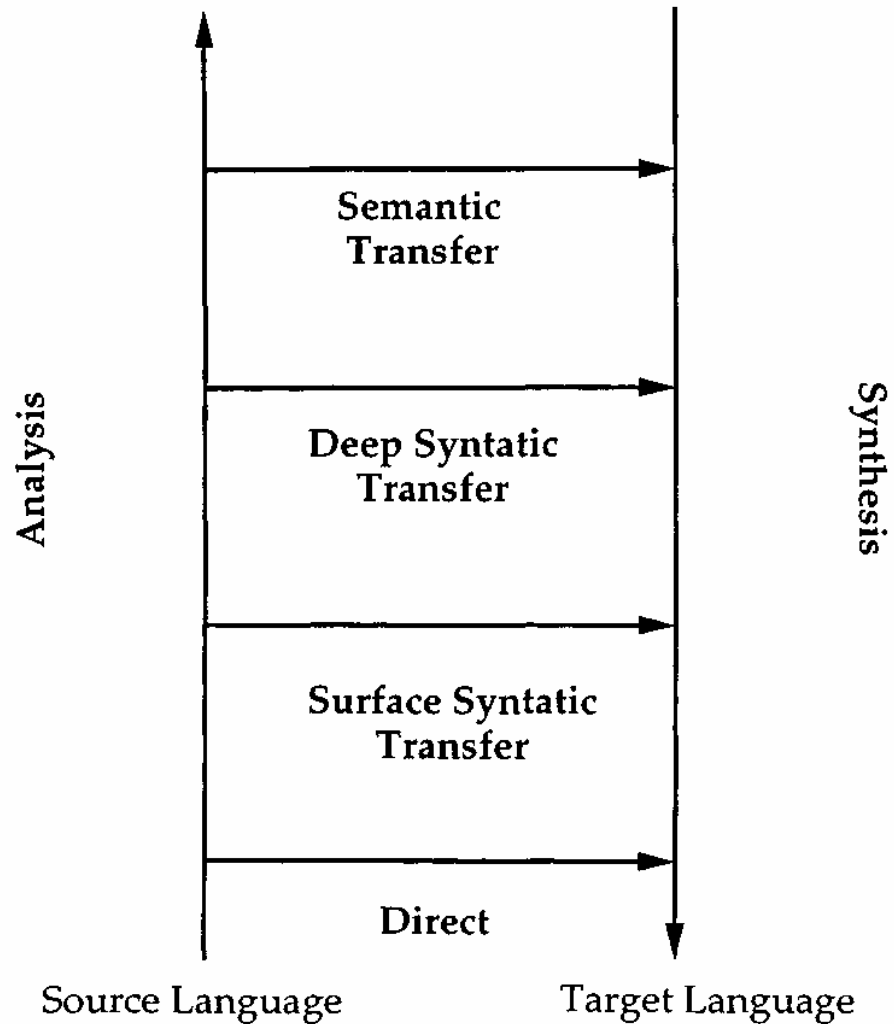
An Interesting Notion :

R&D effort is proportional to path length

An Alternative to the Future
of MT Technology

Analysis

Synthesis

Semantic Transfer

Deep Syntatic Transfer

Surface Syntatic Transfer

Direct

Source Language   Target Language

Some Possiblities :

Transform whenever is appropriate
Apply Machine Learning technology in
narrowing the gaps.