# What should we do next for MT system development?

**Hozumi Tanaka**
Tokyo Institute of Technology

## 1 Introduction

Machine translation (MT) research and development began at the end of 1950's when not only natural language processing (NLP) technology but also linguistic theory was at a primitive level. Given the restricted memory sizes and computing power at that time, MT presented one of the most difficult and challenging research themes of the day. Thus. MT researchers and developers were forgiven when they complained that their poor translation results were due to shortages in memory and computing power. But now, we cannot say such things.

About 40 years have passed since then and computing power and memory capacities have increased dramatically in that time. Many new NLP technologies and linguistic theories have been proposed, based on which MT systems have been developed. Consequently, the scale of MT has grown and many MT systems are available at affordable prices.

However, even though the domain of almost all current MT systems is highly restricted, more improvements are necessary, since their translation results are still unsatisfactory. As before, high quality MT remains a difficult and challenging research theme.

In the 1990's, information networks have diffused throughout the world, enhancing the importance of MT systems. Through the Internet, we find ourselves surrounded by an enormous amount of documents written in many languages, such that we feel the urgent need for multilingual translation systems. MT technology is now considered a key technology in the field of Internet-based information retrieval.

In this paper, after reviewing MT research history focusing on MT technology (not individual MT projects). I discuss what we should do in the next century.

## 2 1960's

As memory and computing power were not only lim-ited but also expensive in the 60's, vocabulary and grammar sizes were very small and MT researchers focused more on demonstrating the possibilities of MT systems through several experimental prototype systems. They recognized the importance of fundamental research as pointed out by the ALPAC report. MT systems at the time made use of only syntactic information, but, in the field of artificial intelligence (AI). the importance of semantic processing was emphasized. Within the confines of toy systems, handcrafted semantic processing was possible. However. MT had to handle broader linguistic phenomena with larger sized vocabularies and grammars.

In the 60's. an epoch-making linguistics theory. Noam Chomsky's standard theory was born, where transformations were so significant that his theory was called Transformational Grammar (TG). His theory created an optimistic view on some MT researchers. An efficient syntactic parsing algorithm called CYK was developed in the early 60's. Taking the breadth-first search strategy. CYK is a bottom-up parsing algorithm based on Chomskified CFGs.

## 3 1970's

### Deep analysis but very small coverage MT

In the 70's. AI researchers were interested in Natural Language Understanding research(Winograd. 1972), one of the main themes in AI at the time. This had repercussions for some MT systems, where, similar to the AI tradition. MT was based around deep analysis with semantic information. Yorick Wilks proposed a method of semantic disambiguation by preferential scoring. which he called "preferential semantics (Wilks, 1973)." As the preferential scores were attributed through human intuition, it was very difficult to extend this approach to practical large-scale systems. His system was in principle rule-based.

### Shallow analysis over a very narrow domain

Contrary to the AI approach, a practical MT system was initially developed in the very narrow domain, such as meteorological forecast news wires in which highly regulated expressions appeared frequently Even though there were no exemplary features to such systems, they were the first systems with practical applications, the success of which encouraged many researchers to look at more ambitious practical MT system projects.

### ATN, Earley and Chart

ATNs with procedural attachments were frequently used in applied NLP systems such as question and answering systems. With procedural attachments, it was possible to carry out semantic processing. The ATN tradition was succeeded by the Definite Clause Grammar(DCG) formalism in the 80's. but in place of procedural attachments, DCG utilized a unification mechanism. ATNs adopted a depth-first search strategy that was inconvenient when selecting the best of many parsing candidates.

To avoid disambiguation problems, a subset of natural language was proposed in order to make NL analysis easier and reduce parsing results. One of the reasons that the sublanguage approach was not widely accepted was that we were so accustomed to using the full set of NL that we can easily become confused as to what is and is not in the NL subset.

In the field of computational linguistics, various parsing algorithms emerged. Earley and Chart algorithms were based on breadth-first searches over CFGs. which did not have to be represented in Chomsky normal form. With minor modifications. Martin Kay's Chart algorithm could be run either bottom-up or top-down. It has been incorporated into many NLP systems in the 80's.

### Lessons

We have learned a lot from the experiences of constructing experimental MT systems:

- Some syntactic parsing algorithms like Chart are extendible to practical levels, but we encounter the problem of generating too many parsing results when the size of the grammar becomes large. We thus need a scoring mechanism to prune off needless parsing results as early as possible. Parsing algorithms with a breadth-first search strategy seems adequate for pruning.
- Semantic-based approaches are very important in the case of semantic disambiguation, but suffer from shortfalls in the extent of knowledge. The volume of knowledge has always been too small to construct a practical MT system. We have to build a large knowledge base (KB) that includes both ontological and lexical knowledge.

## 4 1980's

### Shallow analysis but larger-scaled MT systems

The 80's were the most active period in the history of MT. Particularly in Europe and Japan, several big MT projects were launched. Due to rapid progress in LSI technology, computing power had finally reached a level sufficient to build quite large MT systems. Many projects benefited from the financial support of governments and private companies. Most efforts were aimed at developing practical MT systems aimed at limited domains such as scientific documents, news, or instruction manuals. In the mid 80's. several MT manufacturers announced commercial MT products.

### Practical MT systems

Although MT systems had reached a nearly practical level, commercial MT products in the mid 80's did not meet cost effectiveness criteria: because MT systems were priced highly, customers complained about the translation quality. Systems needed a lot of human intervention, particularly in the pre-editing of source texts or post-editing of target texts.

Analysis of the source text was inevitably shallow due to the insufficient volume of knowledge accumulated to this point. Even though the domain of MT had been restricted, systems had to handle quite a wide range of linguistic phenomena. As a result, the grammar sizes increased and grammars became more complicated, so as to be able to deal with many fringe linguistic phenomena. Semantic grammars introduced semantic categories as non-terminal symbols in CFGs. so as to enable semantic processing in combination with syntactic processing. This approach was successfully adopted in some MT systems aimed at very narrow domains such as stock market news in the next decade.

### Toward deeper analysis

Around the middle of the 80's, some MT projects tried to incorporate complex feature structures into the description of each dictionary entry. This allowed for deeper analysis of the source text. The knowledge acquisition bottleneck described above led researchers to develop large KBs that included concept-level ontologies. Typical KB development projects were the EDR project in Japan and CYC project in US.

### Advances in linguistics

As part of the evolution of his linguistic theory, Chomsky published a new linguistic theory called GB. Discontent with Government and Binding (GB) brought about the birth of other linguistic formalisms such as LFG, HPSG, UG and DCG in this decade. Some theories emphasized the importance of the lexicon and the content of lexical entries, and excluded transformation operations.

Another distinguishing feature of these theories was the utilization of unification, which played a significant role in their implementation. Compared to transformation operations, unification operations had many desirable characteristics from a computational point of view. Therefore, computationally tenable linguistic theories became a hot research theme in not only the theoretical linguistics but also the computational linguistics community. UG, LFG, HPSG and DCG adopted unification as their fundamental operation, making their theories more transparent in a declarative way.

### Chart, GLR and NLP Tools

As I mentioned above, the Chart parsing algorithm was invented at the end of the 70's. In the 80's, in addition to Chart parsing, GLR parsing emerged and grew in popularity in the field of NLP. Interestingly, both Chart and GLR parsing run in a bottom-up fashion with a breadth-first search strategy. Both can usually yield many parses in order of preference if necessary. Although the time complexity of GLR exceeds that of the Chart algorithm, empirical experiments demonstrated that the actual parsing speed of GLR was comparable to that of Chart. Many NLP tools implemented these parsing algorithms.

### Example-based approach

In selecting the correct translation in the target language, it is necessary for us to perform word-sense disambiguation through deeper semantic analysis. Owing to the difficulty of deeper semantic analysis of source texts, the so-called example-based or analogy-based approach was proposed by Makoto Nagao, which works as follows.

Given a sample set of typical translation sentences, each of which is composed of a paired source and target language sentence, we analyze a novel input sentence through comparison with each sample source language sentence, and identify the sentence most similar to it. The translation of the most similar sentence provides strong pointers to a correct translation for the input sentence.

This method seemed to alleviate the overhead of deeper semantic analysis, but it relied on the calculation of similarity between the input sentence and sentences in the example base. To calculate the similarity between two sentences, many researchers found the importance of building a large ontological KB. However, in general, we can expect the translation task to be harder than it would appear.

### Interlingua and the transfer method

MT researchers and developers were engaged in the controversy, "Which is better, interlingua-based or transfer-based MT?" The proponents of the interlingua (IL) method emphasized the merit in constructing multilingual MT systems, the needs of which would grow along with the rapid increase of communications through the Internet. While recognizing the difficulties in designing an ideal IL, they believed the importance of IL-based MT would progressively increase in the near future. On the other hand, some opponents were doubtful of the existence of an IL that was independent of any language. Instead, for more practical reasons, they preferred the transfer method to IL-based MT.

At the end of the 80's. IL-based MT products were announced to have been implemented. However, they were nearly equivalent to the transfer-based MT systems, with the only difference being their slant toward performing deeper analysis of the input sentence. It is true that we do not yet have IL-based MT systems in the strict sense, but the efforts toward IL-based MT should not be neglected in a longer-term view of MT.

## 5 1990's

### Corpus-based and statistics-based approaches

Syntactic parsing is more suited to breadth-first than depth-first methods because the former more readily yields parses in order of preference, enabling us to prune off less preferred parses whenever it is necessary. One of the problems with this approach is the quantitative definition of syntactic preference, and observation that parse preference is closely related to semantic felicity. As semantic scoring remains a difficult task, the preference score is generally calculated in terms of statistics with a strict mathematical founding.

After the success of statistics-based approaches in speech recognition in the 80's, a variety of probabilistic language models have been proposed. Here, a statistical preference is calculated for each parse. Even though the n-gram language model had been successfully applied to speech recognition tasks, it was overly simplistic to be used in the syntactic parsing of natural languages. A more sophisticated probabilistic language model was needed.

The effectiveness of the Probabilistic CFG (PCFG) language model was demonstrated in the 80's. However, as it was a CFG-based model, it was unable to model any context sensitivity. Fortunately, two-level PCFG and Probabilistic GLR (PGLR) language models were proposed in the early 90's, which could naturally incorporate some context sensitivity into the proposed language model, enabling a probabilistic score of preference to be attributed to each parse.

With respect to the PCFG language model, disambiguation experiments empirically demonstrated the need for better corpus-based or statistics-based language models, and large-sized labeled corpora were required to train the probabilistic model. Probabilistic TAG also enables the incorporation of mild context-sensitivity into its language model, and good experimental results have been achieved.

Note that statistics-based approaches contribute to keep the search space narrower for semantic processing.

### Example-based approach, revisited

Example-based MT was proposed in the 80's, but it was not an easy task to compute the similarity between a given input and each example sentence. Furthermore, the speed and quality of translation degrades as the size of the example base increases. The larger the volume of translation examples, the more frequently confusions in similarity calculation occur. This fact is intuitively contradictory, since a human translator will tend to develop greater translation speed and skill as he has exposure to more translation examples. This is the reason why example-based MT technology has been restricted to deal with the analysis of only highly regulated expressions. However, example-based MT seems to have a psychological founding. The pitfalls of the current level of example-based MT are due to a shortfall in learning ability thorough concept level generalization, which is an important research theme for the future.

### Multilingual MT

As mentioned above, there have been a few attempts to build a multilingual MT system. Center of the international cooperation for computerization (CICC) in Japan began a multilingual MT project, covering MT between 5 Asian languages. The project tried to build a multilingual MT system by introducing an IL. However, the project team did not have a satisfactory IL which received the full support of all researchers from the five member Asian countries. In actuality, most MT systems developed thus far have been transfer-based systems targeted at a fixed language pair, in which English was the most frequent choice for a target or a source language.

### Simultaneous interpretation

A simultaneous interpretation project was initiated at ATR in Japan in the mid 80's and still continues today. Simultaneous interpretation requires on-line real-time MT through spontaneous speech understanding. As spontaneous speech contains a lot of noisy ill-formed sentences, robust NLP techniques prove an important research theme yet to be solved.

### Language resources

As mentioned above, corpus-based approaches are unable to solve many intrinsic NLP problems, but are useful for disambiguating parsing results without performing deeper analysis. Increasing the size of training corpora produces a more precise probabilistic language model. Many researchers and developers of MT recognize the importance of linguistic corpora. However, immense human effort is unavoidable in developing a very big linguistic corpus[1]. In the late 80's through to the 90's. MT researchers worked towards developing common LRs that can be shared, incrementally improved and stockpiled.

KB is essential in developing high quality MT systems. WordNet, developed under the direction of George Miller, is one such KB. It is publicly available, and has been utilized in many NLP systems. WordNet will undoubtedly lead to desirable results for future high quality MT efforts based on NLP understanding.

## 6 What should we do in the 21st century?

### MT through NL understanding

Anyone would agree that a good human translator makes full use of semantic information as well as contextual or discourse information when performing translation work. More proficient translators are able to fully comprehend the target text, reorganize it at the concept level, and then produce the target text. There is often no sentence-by-sentence correspondence between the source and target texts, as translation is carried out at the concept level. In the 21st century, we must devote our efforts to constructing an MT system based around NL understanding, even if we know it is not only ambitious but also difficult as a research theme. Without such efforts. MT systems will not be able to escape from a specific domain such as scientific documents or instruction manual texts. Neither MT of literary works nor high quality MT will be possible. Breakthrough for the next generation MT systems will be given birth from NL understanding technology.

There have been great advances in both morphological and syntactic processing technology in the second half of the 20th century. Deeper semantic processing along with discourse processing will be key technologies for NL understanding-based MT systems appearing in the 21st century. Very sophisticated KBs will also be necessary in order to perform deeper sentence analysis incorporating semantic and discourse analysis. Actually, NL understanding might come naturally given such a sophisticated KB. Sophisticated KBs and high quality MT should be developed hand in hand in the 21st century.

We ought to concentrate more on advances in knowledge acquisition technology stemming from AI research. Within the framework of NL understanding, IL-based multilingual MT systems would seem to comprise a real-world realizable application. Also, as NL understanding and KB technology are key technologies of both AI and MT, AI and MT researchers should cooperate more to solve these difficult problems.

[1] Hereafter, we use the term, "language resource (LR)" in a wider sense, which includes linguistic corpora, ontological and lexical KBs, and linguistic tools.

With respect to statistics-based approaches, at least in the short term, we should work towards developing a disambiguation method which makes use of al-l of statistical scores, syntactic parse scores and co-occurrence scores, the latter of which we will be able to calculate from the semantic tags of co-occurring words in a sentence. The method is called a hybrid approach, a combination of rule-based and statistics-based approach.

The following are a selection of issues which remain quite difficult in the 90's and should be solved as early as possible in the 21st century:

- identification of coordinate structures,
- dependency analysis of long sentences.
- analysis of spontaneous speech and its translation
- handling of ill-formed sentences.
- realtime NLP with limited computing resources.
- word sense disambiguation.
- ellipsis resolution.
- identification of anaphora/cataphora.
- generation of NL with wide coverage.
- developing a large scale KR and KB.
- design of IL.
- multilingual MT.

## Language Resources

We have elucidated that languages resources (LRs) are, in a sense, the infrastructure of NLP, and the effectiveness of LRs in the field of NLP has been demonstrated in the 90's. In addition to NLP, LRs are useful not only for enumerating linguistic phenomena, but also evaluating NLP systems. LRs are very important for linguists and NLP researchers including MT researchers.

Experiments in the past have shown that, the larger the volume of LRs, the better the quality of NLP. However, we cannot rely solely on human labor to develop extensive LRs with complex annotation, as this is not only tedious and time consuming but also calls for massive human resources.

To solve the above problems, it is natural to say that NLP technology might be helpful in building either complex or large LRs. This would take the form of a kind of a bootstrap method, an interesting research theme which should be tried out seriously in the next century. As LRs include ontological and lexical knowledge, they have strong connections to knowledge acquisition/discovery technologies developed in AI. In this respect, MT researchers should be more aware of what is happening in the field of AI.

As regards MT. I would like to point to the necessity for greater efforts to construct multilingual or parallel corpora.

## MT and the Internet

A tremendous amount of text, written in a wide array of languages, already exists on the Internet. With the advent of a large-scale Internet society, we very often encounter the need for translingual information retrieval, which is a very active research area. As it is impossible to build better translingual information retrieval systems without the aid of MT technology. more investment should be made in MT. Needless to say. multilingual MT systems are the best choice for this purpose. We should not forget that MT will enhance our Internet society in the 21st century.

According to the book “The future of English” by David Graddol (Graddol, 1997), in addition to the globalization of economic activity and science, high-technology, and particularly computers and the Internet, have accentuated the spread of usage of English throughout the world. However, the claim is that by the middle of the 21th century, even if English continues to dominate other languages, her influence will be on the decline. There are more than 6,700 languages in the world. 33% in Asia and 19% in the Pacific, which means that more than 50% of the world's languages come from Asia and the Pacific. It is possible to conclude that in the 21st century, languages used in Asia and the Pacific will become more and more important, in tandem with economic growth and advances in science and technology in this region. The need for MT systems bridging the gap between these languages, will be felt more keenly as a result. The following ranking of anticipated mother tongue populations in 2050 along with 1996[2] elucidates this fact:

| | | | |
|---|---|---|---|
| 1. | Chinese | (1.384 billion in 2050) | [1.113 billion in 1996] |
| 2. | Hindi/Urdu | (0.556) | [0.316] |
| 3. | English | (0.508) | [0.372] |
| 4. | Spanish | (0.486) | [0.304] |
| 5. | Arabic | (0.482) | [0.201] |
| 6. | Portuguese | (0.248) | [0.165] |
| 7. | Bengali | (0.229) | [0.125] |
| 8. | Russian | (0.132) | [0.155] |
| 9. | Japanese | (0.108) | [0.123] |
| 10. | German | (0.091) | [0.102] |
| 11. | Malay | (0.080) | [0.047] |
| 12. | French | (0.076) | [0.070] |

This ranking suggests that Chinese, Hindi, English. Spanish and Arabic will still keep the top major languages in 2050.

[2] By the engco model

It is also expected that about 90% of the world's languages will become extinct in the 21st century. We are not sure what kind of influence MT will have on this problem. It seems an interesting question to ask whether or not MT will have a positive or a negative effect on the survival of minor languages. Translating major languages has been a traditional goal of MT, justified from an economic standing and not for cultural reasons. My personal opinion is that MT researchers should pay more attention to the minor languages, based on a re-evaluation of the future of minor languages in the 21st century.

## Standardization

Most languages have their own character sets, which causes a lot of troubles when using computer systems. Some of them do not even have an internationally recognized character code set. Because of incompatibilities between code sets, especially in developing countries, computer systems physically connected over the Internet cannot communicate with each other in their native tongue.

I have already mentioned that more than 50% of the world's languages are spoken in Asia and the Pacific. The MLIT project, sponsored by CICC[3] in Japan, aims to propose a standard character code set for each Asian country, with the cooperation of the following countries: P.R. of China, HK SAR, India, Indonesia, Japan. Laos. Malaysia, Mongolia, Myanmar, Nepal, the Philippines, Singapore. South Korea. Sri Lanka. Taiwan R.O.C., Thailand and Vietnam. The MLIT results shall be collated as a proposal to ISO early into the next century, which is hoped to contribute to multilingual MT system development.

As well as computerized character code sets, the standardization of tag sets in LRs is also very important. For instance, the standardization of a part of speech, one of tag sets, is possible with only considering one language.

On the other hand, the ontological knowledge in LRs seems different from culture to culture, and the standardization seems a difficult task. The upper concept level, however, seems to have much in common and be language and culture independent. If it is possible to achieve this standardization, it will be immediately applicable to the design of an IL.

## Miscellaneous points

I did not mention the importance of human-aided MT using translator work benches to supplement the weaknesses of current MT systems. There are many interesting research themes related to this, such as the development of better human interfaces, segmentation of long sentences into component sentences, and preediting and post editing technologies. Finding better evaluation methodologies is another research topic. Drawing on the experiences of expert systems like empty MYSIN, empty MT systems should be implemented, to act as a bare engine independent of any domain or language.

Finally, I would like to conclude by stressing the importance of rewriting some MT systems. As MT systems become very big. there is a tendency for no one person to comprehend the MT system in its entirety. This situation is similar to that of the software crisis illustrated by the book "Mythical Man Month (Brooks, 1995)". I recommend rewriting whole MT programs in order to make systems more compact, more transparent, and easier to understand and extend. As the commercial MT products are now available at greatly reduced prices, the customer seems to refrain from complaining about the quality of MT systems he is using. He should give MT developers more severe criticisms. which are very valuable for MT developers to build a better MT systems.

## References

Brooks, F. P. 1995. *The Mythical Man-Month.* Addison Wesley.

Graddol, D. 1997. *The Future of English?* The British Council.

Wilks, Y. 1973. An artificial intelligence approach to machine translation. In R. C. Schank and K. M. Colby, editors. *Computer Models of Thought and Language.* W. H. Freeman and Company, pages 114-151.

Winograd. T. 1972. *Understanding Natural Language.* Academic Press, New York.

[3] http://www.cicc.or.jp/
Committee on Multilingual Information Technology. *Joint Development Research on International Standardization: Multilingual Information Technology.* 3 1999.
Please mail info@net.cicc.or.jp for more details.