3

Artificial Intelligence Contributing toward Nation Building

Professor Dr. Zaharin Yusoff
President, Multimedia University
Cyberjaya, Selangor
Malaysia

I will be talking on work which is "Qualitatively Quantitative" but my behaviour is normally "Quantitatively Qualitative". That means yakking a lot. Essentially, I'll be telling a story about results of research which is in *machine translation,* a domain within natural language processing which is part of artificial intelligence.

Computer Science and Information Technology

The story-the case study is about our struggles, about how we worked to get as close as possible to the market and the problems we faced. To begin with, as I mentioned earlier, we will be talking about computer science and within computer science is artificial intelligence. The work I will be talking about will be in language technology.

So, if we are talking about contributions to nation building, I think it's very clear computer science-IT contributes toward nation building because it looks for applications. The ultimate results are systems and applications that contribute to devices that would simplify, improve efficiency, productivity, at work or life in general.

Sub-Domains and Artificial Intelligence

Some of the domains are *information systems* is for organizing information, *networks* is for connecting people together, *high performance computing* make them go faster, *software engineering* to do it properly and *artificial intelligence,* which is my domain is to make them less stupid.

So, artificial intelligence is mainly about intelligence. Intelligence means being able to perceive its environment and takes actions which maximize its chances of success. The domain is about science and engineering of making intelligent machines. So, these are some of the areas that qualify to be in artificial intelligence (deduction, reasoning, problem solving, cybernetics and brain simulation, search and optimization and languages). Actually, Cognitive sciences overlap with artificial intelligence and where I am heading is towards natural language processing.

Artificial Intelligence

Essentially artificial intelligence hopes to compliment or maybe replace humans with machine with some level of intelligence, could remember everything and does not tire. That's the goal. The domain that we work in is sometimes called language technology, is known as natural language processing or computational linguistics.

This is my definition of artificial intelligence; "*The understanding and explication of language phenomena in a computationally tractable form resulting in techniques for interchanging various linguistic forms (speech, text, morphology, syntax, semantics/meaning, discourse, knowledge), thus leading to the creation and development of intelligent applications involving language*". Essentially it is anything to do with language.

Within that, we notice language and knowledge are two common things to all humans. Language is an expression of knowledge, and knowledge is the essence of intelligence and cuts through different languages and various modes of communication. That's a mouth full but essentially if we can handle knowledge and language then there no reason why we cannot build applications that are multi-lingual, multimodal and also intelligent. So, within this lies the translation technology.

Translation Technology

In translation and modes of communication, we also look at translation from images to language. It is essentially comes from both ways, from *speech* to *text* all the way finally to *knowledge*[1].

Meaning somewhere in there (within the process) so the work is going down-there (speech) to up-there (knowledge) or from up-there (knowledge) to down-there (speech) and they will be based on whatever applications.

Translation from *text* to *meaning*, we do it in one language. From the *meaning*, because meaning is independent of language, we generate it to go backwards in another language. That's essentially what translation is. These are some

[1] Diagramic Depiction: Speech, text, morphology, syntax, meaning, discourse, knowledge – all in a sequence and flow up and down with speech being at the bottom.

examples of applications in language technology; *Speech (text to speech/ speech to text), Multilingualism (translation, multilingual dictionaries), Meaning Based (document categorization, information retrieval) and Knowledge Management (identification, acquisition).*

The Case Study in Translation Technology

The origin

I was at USM (Universiti Sains Malaysia) for 25 years. In fact machine translation started before I even joined USM. In 1997 there were only USM partnering with GETA from France, that's where I went and later joined by UTMK (*Unit Terjemahan melalui Komputer*). From there many new players came from UTM (Universiti Teknologi Malaysia), UKM (Universiti Kebangsaan Malaysia), UM (Universiti Malaya), UiTM (Universiti Teknologi Mara) and so on.

Types of work then in this field

And then essentially everyone was working *on computational linguistic tools* and also on *collection of data*. That is the two types of work that one does and it was from these we developed applications. So, translation would be the tool for example as an analyzer for English language and the generator for Malay language and that we used data-mining dictionary and so on. Although at that time we had many players, the number of researchers where not too many.

The resulting work

The following tools are the result within those years before 2000; Generic Tools (Spellchecker, Text Analysis), Application Based Tools (WEB crawler, portal generator) and Lingware Data (English to Malay, Malay Grammar). Then data, those days because of space problem it was difficult to collect data.

2001 to 2005

After 2000, this is after the CISCC project with the Japanese, French project (and) many (more) projects but everyone was trying to go for the fully automated translation, getting the machine to do everything. So what happened then from 2001 and 2005 was this...nothing! We just could not get grants. We just could not do whatever. We were hiding behind somewhere still working on it but hiding the project behind other practical more commercialise able projects.

The revival

Then, came the revival about in 2005; this was the extension of the 8th Malaysian Plan. If you remember, the money came late for the 9th Malaysian

Plan, there was an extension, extra time with a little bit of money but no golden goose. By then, we had come out of the shadows and we were financed by the part-time grants. We were working with several universities locally and internationally.

Later, I went to Malaysian Institute of Microelectronic Systems (MIMOS) because there was a promise of a large chunk of money to do language technology. I was seconded from USM and USM *(for example, Computer-Aided Translation, Grammar Formalisms, and Dictionary Processing)* had all these tools, UKM *(for example, Information Retrieval, Lexicography, Morphological and Cognitive Linguistics)* had all these tools and so on. Dewan Bahasa dan Pustaka (DBP) came into the picture, we had MIMOS, we had *Standards for Character Code for Jawi Word Processor* (SIRIM) and all put together, working together on translation and whatever else we could do in language technology.

From 2000 onwards with so little data because this domain requires lots of data, here and there, these was this major effort in collecting of dictionaries, thesaurus, and bilingual phrases and so on. Having all of these, we were then consolidating it.

Deployment of translation service – An attempt

What we wanted to do was this - deploys translation services because there was no use in trying to get machines to translate on its own. It was too difficult. So, we were combining Computer Aided Translation (CAT) workbench where the computer did the translation aided by human. Here, we had the dictionaries, terminologies and thesauri. Worst comes to worst, it was the human that does the translation.

There was also a Translation Memory whereby previous translations could be referred by the machine to determine what the translation should be.

The translation was carried out by someone sitting in front of a CAT workbench which was one of the applications. A person could, for example make use of the machine translation to provide a draft and then he/she could edit it. If the draft comes out rubbish, he/she could just make use of the Translation Memory to help out. The time then, was right to move for services over the Internet and also through Short Message Service (SMS).

Derailed

Well, that was the idea and we almost got there. Unfortunately, there was a change in direction at MIMOS. This project was derailed even though many universities were keen.

We failed the first time before 2000 because we were trying to achieve too much, trying to get the machine to translate by itself. Later, being realistic, we were trying to combine human and machine translation. This was actually very feasible. Unfortunately, the funding ran out.

2006 to 2010 Let's Try it Again

So, from 2006 to 2010 which is now, let us try it again! Actually, we took a breather from 2007 and started again in 2008. We were working anyway even when the funding stopped. We could not get enough funding to put it together but still we were able to work on bits and pieces.

When we resurfaced, by then we had new grants: Science Funds, Techno Funds and Malaysian Technology Development Corporation (MTDC). We are now close to the product. The message is that as much as things may appear as if they failed, work can still go on. Then, we can still move up the scale[2].

This Time Around

This time there is something called the *Snake.* So, the idea here is not only do we look at the translator or someone using various machine like translation memories and what not to translate but also using people to gather the data for us. There were expert teams from various universities using *Snake* to upload existing terminology bases and translation memory to build the *Bi-Lexicon* which was the core for *Bilingual Knowledge Base* (BKB).

We would have expert teams to build the database. You also use this system to gather parallel text and then, we bring in the translators. When they start translating, they will upload their own data because each has personalized dictionary and so on. Now, we have a system which actually learns and picks up data as it goes along.

Changing the Translation Methodology

We are also trying to change the way translation being done. The conventional translation approach is that you have *professional translator,* then, you give them the *source text* and they do whatever they want with it and get the t*arget text*.

Later, we introduced Just to Translate (J2T). Using J2T methodology, the *source text* goes though a quick *rough translation*. Then, we get what we called the *Example-Based Machine Translation* (EBMT) text.

[2] The *Scale* here refers to the flow from concept, proof of concept, laboratory prototypes, industrial prototype, product to commercial product and services.

And then, we have a small group of *just translators*, anyone who uses the tools for editing. This is done offline and sometimes or more often the translation is good enough. For publication materials, we will bring in the *professional translators*.

This is done offline and sometimes or more often the translation is good enough. For publication materials, we will bring in the *professional translators*.

So, we worked out that the usual translator can translate about 1,250 words a day. But, using our method, we can actually work much faster and the cost also differs. So, using J2T approach, we had a 25% reduction and a faster speed.

So, it has evolved from a little black box where you shove in text and then you hope for the best for what comes out. Now, it is getting more sophisticated and it has ways and means of taking full advantage of the participation of human and also the availability of data.

Privatization

Looking at our situation now and also as a reminder of certain things you see, technology is one thing but getting it to the market is not that easy. This is what we call privatization. In fact, we as universities normally want to just jump into commercialisation. It was very difficult! We really had to *productise* it especially in Information Technology (IT). Then, we also had to have support from companies because university lecturers are not capable of doing so. The companies would actually distribute the service products and act as solution providers. Only then, can we get into the market.

It is nice to say a researcher can go out there and sell. Now, you have to be an entrepreneur to do that! It is better to let the researcher stay where they are. We will find the right people to do things after (privatization).

As we go along we have to lose a bit of the ownership of the intellectual property, we have to pass it on to others in order for them to market it.