IT2025 Supplementary Note 03

October 22, 2025

1 Word Embedding

As mentioned in the lecture, all large language models are computational models. They only accept number inputs and then generate number outputs. To input a sequence of words to a large language model, one pre-processing step is to convert each word or phrase to *an array of multiple numbers* (i.e. a vector) via the method of word embedding. After the large language model has generated a sequence of vectors, these vectors are then converted to a sequence of words or phrase as output. The idea is illustrated in Figure 1.

1.1 Language-Oriented

Note that the result of word embedding relies on a bag of words and the use of language in document writing. If the bag consists of English words only, the word embedding fits for English language. If the bag consists of Chinese words only, the word embedding fits for Chinese language.

As use of a language might be different from one language to another, a *Chinese prompt* to a LLM being trained with *an English word embedding method together with a translator* might give different output from the LLM being trained with a Chinese word embedding method alone.

1.2 Token

A token refers to a word or phrase which can be encoded by a single vector. A vector is an array of multiple numbers. For instance, a location of an airplane can be represented by three numbers usually called (x, y, z), which is an array consisting of three numbers. In mathematical term, it is called a three-dimension vector.

So, the size of an input prompt (resp. output response) is normally counted by the number of token instead of the total number of words.

2 Factors Affecting the Performance of a LLM

From the above discussion, one should note that the performance of a LLM depends on at least four factors.

- Methods of word embedding for the LLM.
- The computational model in the LLM.
- Methods of word embedding for the translator.
- The computational model in the translator.

It should be noted that an AI translator is yet another computational model, Figure 2. An application developer should be aware of these issues.

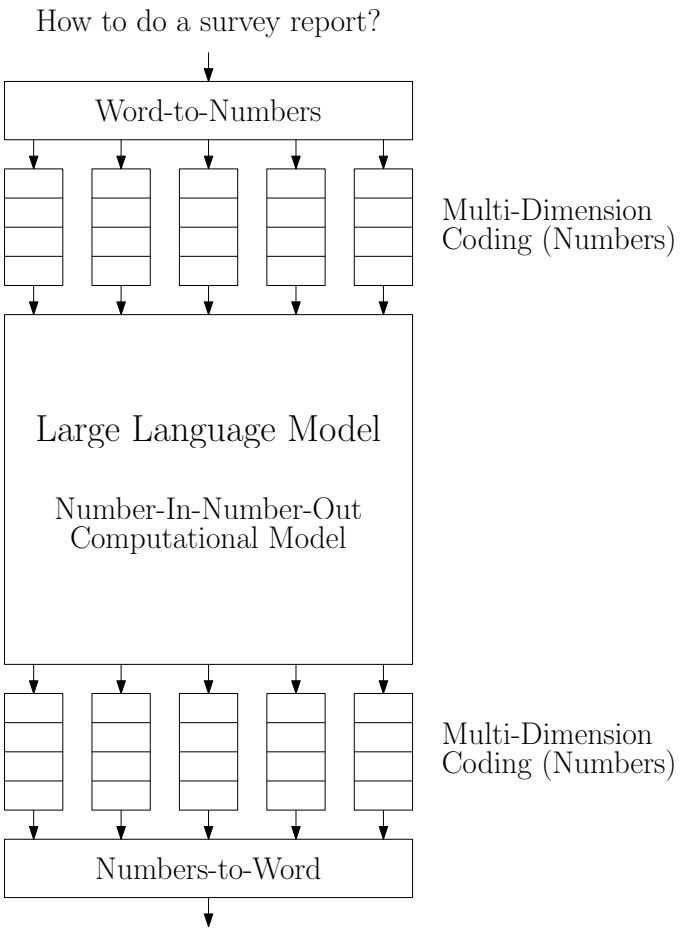

Figure 1: Illustration of the function of word embedding. For an input prompt, it is used for converting each word or phrase from the input prompt to a multi-dimensional vector. Once the LLM has generated a sequence of vectors, it is used for converting each vector to its corresponding word or phrase. This sequence of words and/or phrases will be output to the user in response to the prompt.

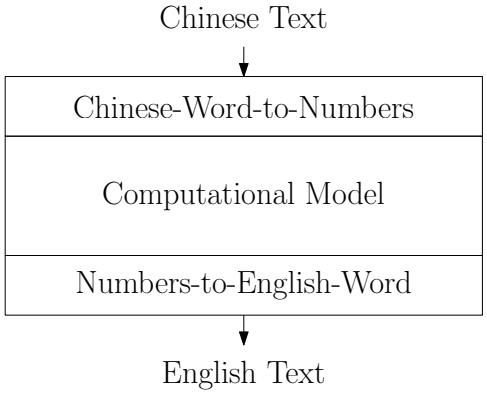

Figure 2: An AI translator is yet another computational model. Besides, it consists of a *Chinese-Word-to-Numbers* module and a *Numbers-to-English-Word* module.

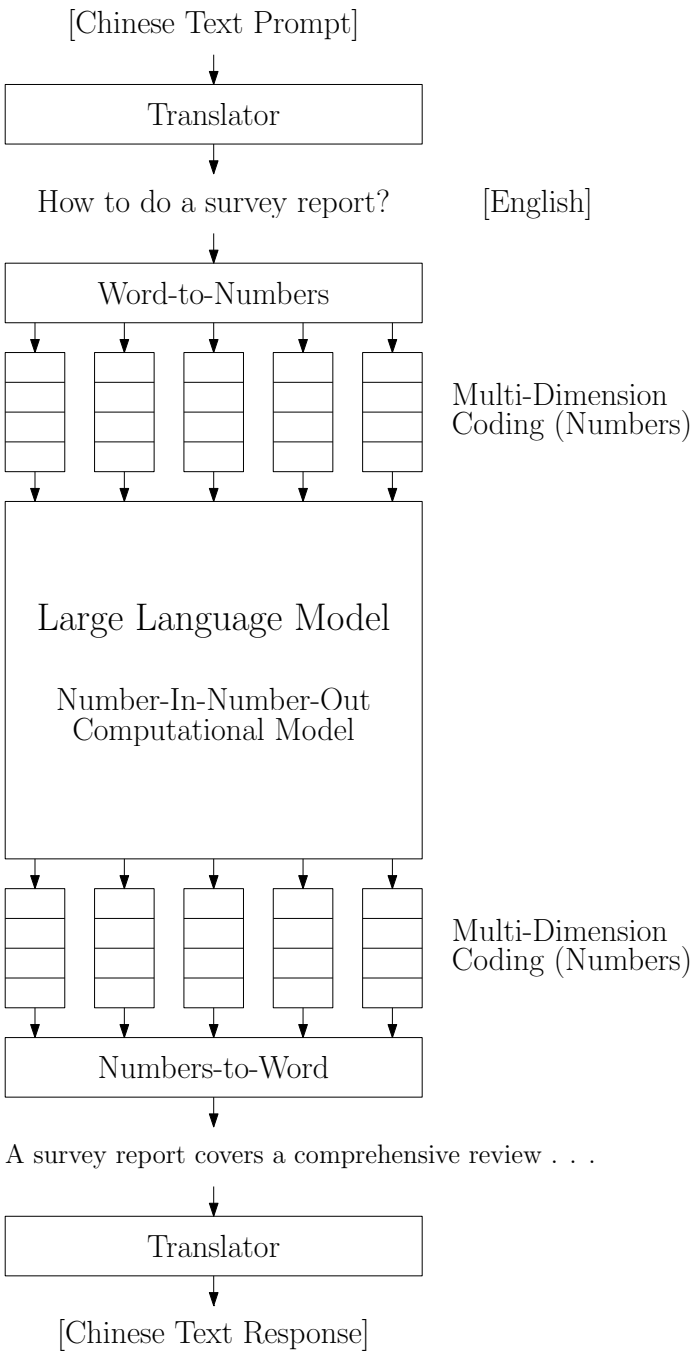

Figure 3: Illustration of the function of word embedding and translator in prompt-to-text generation. It should be noted that a translator is yet another computational model. Word embedding systems are embedded in the translators as well. For the input prompt part, Chinese-Word-to-Numbers module and Numbers-to-English-Word module are embedded. For the output response part, English-Word-to-Numbers module and Numbers-to-Chinese-Word module are embedded.