Author's contact:
Lifeng Han lifeng.han@adaptcentre.ie Tell: +353-892221631 (Postal: 3 Shanliss Drive, Santry, Dublin 9)

Author's supervisors:
Prof. Alan Smeaton (alan.smeaton@dcu.ie) and Prof. Gareth Jones (gareth.jones@dcu.ie)

Affiliation:
DCU, Faculty of engineering and computing, school of computing, ADAPT Research Centre

(*Topics*: Language: Natural Language Processing, Sentimental Analysis, Machine Translation)

Paper title:
Unveil the myth of Chinese characters in neural machine translation

Abstract text:

Background:
Machine Translation (MT) researchers find it hard to achieve both fluency and adequacy levels of the model translation output. For Chinese MT, there has been a lack of work about character decompositions, thought such kind of decomposition knowledge was successfully applied in other fields, such as sentiment analysis and classifications. Especially, the comparisons of different decomposition levels and their investigation in MT performances. Chinese is an evolution of hieroglyph and partials of the characters such as radicals and strokes carry a significant amount of information that deserves a deep analysis.

Objective:
Investigate the performance of Chinese character decomposition for MT and linguistic aware MT.

Methods:
The state of the art Neural MT (NMT) with attention mechanisms applied on encoder and decoder is set as the baseline. The proposed model applies Chinese character decomposition into three different levels and testifies which level represents best the original character sequence in meaning. Multi-word expression (MWE) terms as linguistic knowledge are used to enhance MT.

Results:
Experiments show that, interestingly, levels one and three achieved better representations than level two; the level three decomposition which is cipher-text for native speakers can be surprisingly well learned by NMT models to acquire the original text knowledge. The extracted bilingual MWE further improved model performance score which won the start-of-the-art baseline.

Conclusion:
Efforts are made to unveil the myth of Chinese characters in NMT by decomposing them into pieces of radial and strokes. Better evaluation scores are even achieved with linguistic aware NMT models using MWEs.