Abs-130

Suhaila Saee, Ranaivo-Malançon Bali, and Tang Enya Kong (Multimedia University, Malaysia)

The construction of computational morphology resources for U-RL

Computational Morphology Resources are resources needed for morphology analysis and generation in Computational Morphology field. In Asia, only a small number of listed languages has started to do research on morphology analysis and generation. In fact, Sarawak languages are Asian languages excluded from the list. To our knowledge, there does not exist any applications in the Natural Language Processing context which involve Sarawak languages. For example, morphology analysis and generation that is known as a first process in text processing area. In addition, Sarawak ethnic languages could be categorized as Under-Resourced Languages due to the lack or no digitized resources yet in the Information and Communications Technology context, in general and traditional technology as well.

Morphological data acquisition is a crucial and hard step in pre-processing stage of automatic morphology analysis and generation due to lack of linguistic resources in terms of lexicon, corpus, and grammar. Therefore, in this paper, we highlighted our main problem which is no resources at all in term of unavailability of internal structure when word is analysed in the case of Under-Resourced Languages. The issue related to this is how to get the required resources? There are two types of most required resources in automatic morphological system which are i) corpus and ii) list of stems and affixes. Both resources play an important role to the next steps in morphology analysis and generation, indeed.

On top of that, two research questions have been encountered from the mentioned problem and these are: i) What is the best work flow for corpus acquisition by considering the Under-Resourced Languages issues? and ii) How many data sets are needed to acquire morphology information in morphology induction for Under-Resourced Languages?

In this paper, a work flow for corpus acquisition in the context of Under-Resourced Languages has been proposed. The workflow consists of three main stages which are: i) Stage 1: data collection [three types of sources are dictionaries, grammar book(s), written text], ii) Stage 2: text formation [two possible processes would involve either digitisation (mainly for hardcopy version of sources) or conversion (mainly for softcopy version of sources)] and iii) Stage 3: compilation [compiling three sources that are in the text format into one text file to produce unannotated corpus]. In fact, the three stages are depending on each other in order to construct the corpus. Besides, a result gained from the morphology induction that would be a list of stems and affixes from a very small quantity of Under-Resourced Languages resources also be a part of the proposed solution in this research.

Furthermore, the contributions from this research would be: i) An unannotated corpus that can be used in other Natural Language Processing applications, mainly for Sarawak languages and ii) Morphology information of Sarawak languages that induced from unsupervised machine learning. Last but not least, the challenges from this research would be discussed in the last section of the paper.

References

1. Karagol-Ayan, B. (2007). *Resource generation from structured documents for low-density languages.* (Doctoral dissertation, University of Maryland, College Park). Retrieved from http://www.lib.umd.edu/drum/handle/1903/7580

2. Feldman, A. (2006). *Portable language technology: A resource-light approach to morpho-syntactic tagging.* (Doctoral dissertation, The Ohio State University). Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.615&rep=rep1&type=pdf

3. Cucerzan, S., & Yarowsky, D. (2002). Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th conference on Natural language learning - COLING-02,* pp. 1-7. Morristown, NJ, USA: Association for Computational Linguistics. doi: 10.3115/1118853.1118859

4. Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics, 27(2),* 153–198. MIT Press. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300490