Computational Simulations for Syllable-Based Acquisition of Word-Like Units

In this paper we report the recently obtained results of a set of computational experiments that simulate a syllable-based method for acquiring 'word-like units' from syllabified raw texts, and discuss its implications for humans' acquisition of unknown words. This method is entirely statistical and unspervised, and thus can be considered to represent a domain-general manner of self-learning. The procedure can start from scratch, i.e. where no word is assumed to have been learnt, or from where some vocabulary has been acquired, and thus function as a word learning model at any learning stage.

The proposed learning mechanism targets whatever sequences of syllables of *any length* that are strongly associated statistically, and thus does *not* discriminate between individual words and multi-word expressions (hence the choice of the term 'word-like units', see [2]). Therefore the procedure returns a mixture of words and multi-word expressions, be it named entities such as Barak Obama (or 'ba-ra-ko-ba-ma') or frequent greetings like 'what's up', and this corresponds to the observation that at initial stage of learning children seem to capture linguistic units 'wholistically,' that is whatever sequence of words frequently exposed to themselves, without analysing it.

We use as the association measure an information-theoretic metric, *frequency-weighted mutual dependency*,[4, 5] a variant of pointwise mutual information scores standardly used in collocation detection. Starting from the phonemecised and roughly syllabified text, a procedure based on *variable length N-gram*[3] is applied iteratively from unigram to N-gram, recording association scores of sequences higher than a given threshold, until no sequence is found above this threshold. This procedure gives the highest score to the most frequent, and most co-ocurring sequences. 'Aun San Su Kyi' and 'status quo' are amongst the found examples, whereas non-frequent sequences or, sequences frequent but internal cohesion (say 'of the') will be excluded.

Taking as samples the Twitter data of four languages, English, Vietnamese, Japanese and Chinese, we evaluate the performance in three different manners. First we show the acquisition rate, i.e. how many syllables have been matched to word-like expressions, to see the effectiveness of the procedure. We then compare the results with other competing models of different levels of granularity, first with a phoneme-based model and then for multi-word expressions, with a word-level model. We show that our syllable model is superior in capturing words over the phoneme-level model by a significantly larger margin than the expected difference, and is, for multi-word expressions, not far behind the word-based model. We also compare its quality as a language model, comparing the trigram model that uses the obtained results against the customary word-level trigram model, by means of perplexity, a standard measure of 'unpredictability'. Our syllable model demonstrates

not just an on-par, but superior, predictive capacity for some languages.

A major implication of these results is that they point to the cognitive plausibility, and consequently the potential usefulness, of a syllable-based model for self-learning in vocabulary acquisition. Generally, one of the first tasks for a learner who is exposed to raw sound input is to find words in the continuous stream, and if, as claimed, if syllables are the most perceptible unit on the human ear [1], s/he could intuitively build hypotheses on word segmentations on this basis. Although our main focus has been on first language acquisition, the model can simulate any stage of learning and hence is amenable to extension to such situations. For example, it can be modified to semi-supervised learning, where the student builds hypotheses and the teacher verifies/corrects them. At any rate, further studies are warranted to confirm if the human learner actually follows such a learning path, that is, whether our statistical machine learning model coincides with human cognition, and what stage of learning such a method may be the most effective.

Another important observation is that the performance varies across the languages. In our four target languages, (tone-marked) Chinese reaches consistently the highest score, while English and Japanese generally lag behind. This can be attributed to the pervasiveness of word-bounary ambiguities for English, and to the relatively large number of syllables per word for Japanese. This indicates that the effectiveness of syllable-based acquisition also varies from language to language, calling for a need of cross-linguistic studies. One immediate future plan is to extend the application of our procedure to French, an interesting language with pervasive liaison/elision phenomena. The choice of our current four languages now is based on practicality: a relatively simple (sonority-based) syllabification method works well enough for English, and the phonotactics of the other four are very simple. We plan therefore to build a more sophisticated syllabifier, and given this, a fuller scale cross-linguistic comparison will be made possible, which we believe is of great interest to the research community.

References

[1] J Bertoncini and J Mehler. Syllables as units in infant speech perception. *Infant Behaviour and Development*, 4:247–60, 1981.

[2] A D Endress and J Mehler. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, 2009.

[3] Fabio Kepler, Sergio Mergen, and Cleo Billa. Simple variable length N-grams for probabilistic automata learning. In *JMLR Workshop and Conference Proceedings 21*. Journal of Machine Learning Research, 2012.

[4] Pavel Pecina. *Lexical Association Measures: Collocation Extraction*. Institute of Formal and Applied Linguistics, Charles University of Prague, 2009.

[5] Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of collocation extraction metrics. In *IN PROCEEDINGS OF THE 3RD LANGUAGE RESOURCES EVALUATION CONFERENCE*, pages 620–625, 2002.