LING5700 Discovery Procedure

Problem Set 2: Word segmentation

Word segmentation from continuous speech is the prerequisite for almost everything in language acquisition and has received considerable attention over the years. The prominence of this topic was amplified by a landmark study (Saffran et al. 1996) which shows that statistical learning via transitional probabilities (TP) may be involved in word segmentation. The TP approach, which is described in that paper and many others, is one of the first mechanisms proposed for distributional learning. Harris (1955) suggested that morphemes may be identified as such by applying a TP based method to sequences of phonemes and Chomsky (1955) adopted it for word segmentation even though he did note that the idea had to be verified against some realistic corpus.

Some of us have been critical about the role of statistical learning in word segmentation. The skepticism comes in several independent but interconnected ways: (a) Does TP-based learning work on naturalistic languages (Yang 2004), (b) Are there computational limitations of TP tracking even in artificial languages (Johnson and Tyler 2010), (c) Do humans use a purely statistical method such as TP when other, linguistic, cues, are available (Shukla et al. 2011)? The problem of word segmentation has also generated a large body of computational models, too many to enumerate here. These models, which implement various proposed strategies for segmentation, can be put to test against empirical data.

The paper by Frank et al. (2013) was partly a response to these skepticism. These authors designed a large artificial languages and presented to several adult learners for training on a daily basis and tested them afterwards. The subjects were clearly able to extract, and memorize, some words. Evidently, word segmentation scales.

The question is how. These authors lean toward an account based on statistical learning but did not commit to that interpretation. This problem set investigates. Your task has several components.

- Recreate the stimulus language based on the Frank et al. (2013) paper. It contains enough details so that should be doable. This way you will generate a corpus of this artificial language.

- Your task is to see what kind of word segmentation strategies perform appropriately on the generated corpus. They are two basic classes of models: (a) statistical models that large build around the TP idea, and (b) algebraic models that make use of the factand really an obvious intuitionthat we can chop off previously segmented words in a continuous speech stream to learn the remainder (Bortfeld et al. 2005, Gambell and Yang 2005, Lignos 2011).

- You can either implement the TP learning model yourself or visit

  `https://zenodo.org/record/1471532`

  for a previous implementation.[1] The software package implements several learning models. For the purpose of this project, you only need to run the basic TP learning over syllables.

[1] Caroline Beech used the package for the final project when this class was offered in 2020; the results have been published as Beech and Swingley (2023), which can be found on the class webpage.

- For the algebraic model, you should (probably) visit

  `https://github.com/ConstantineLignos/WordSegmentation`

  for the implemented model of Lignos (2011). Lignos's model has several components. The core is the algebraic method dubbed the subtractive algorithm. One optional component makes use of stress information which some have argued to be informative (e.g., Yang 2004). Running speech, however, rarely exhibits clear, dictionary-like, stress information. It turns out that the contribution of stress to segmentation is quite minimal above and beyond the subtractive method. You should only use the subtractive method here.

- Describe your findings. One goal, of course, is to see which method works the best, e.g., have the highest score of segmentation, but a more appropriate one would be to see to what extent the models match the human behavioral results reported by Frank et al. (2013).

References

Beech, C. and Swingley, D. (2023). Consequences of phonological variation for algorithmic word segmentation. *Cognition*, 235:105401.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., and Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4):298–304.

Chomsky, N. (1955). The logical structure of linguistic theory. Manuscript, Harvard University.

Frank, M. C., Tenenbaum, J. B., and Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PloS one*, 8(1):e52500.

Gambell, T. and Yang, C. (2005). Mechanisms and constraints in word segmentation. Ms., Yale University.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31(2):190–222.

Johnson, E. K. and Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental science*, 13(2):339–345.

Lignos, C. (2011). Modeling infant word segmentation. In *Proceedings of the 15th Conference on Computational Language Learning*, pages 28–38.

Saffran, J. R., Aslin, R. N., and Newport, E. (1996). Statistical learning by by 8-month-old infants. *Science*, 274:1926–1928.

Shukla, M., White, K. S., and Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of the Sciences*, 108(?):6038–6043.

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–456.