Unsupervised learning of linguistic significance

Eric Atwell*[1], Bayan Abu Shawar[2], Andrew Roberts[3] & Latifa Al-Sulaiti[1]

[1] University of Leeds
[2] Arab Open University
[3] Pearson Education

According to the EU PASCAL research network website, "unsupervised learning" means that "the program cannot be explicitly given a training file containing "example answers", and nor can example answers be hard-coded into the program." (PASCAL, 2005). By this definition, our use of machine learning to train conversational dialogue agents using a natural language Corpus (Abu Shawar and Atwell, 2005) qualifies as unsupervised learning: our program does learn from a Corpus, but this "training file" does not explicitly specify example answers to questions the conversational agent may have to handle. The Corpus is merely a set of examples of human conversation; our statistical learning model uses these to infer significant patterns or clusters of potential user input to match templates for system responses. Another example of unsupervised learning is the word-cluster function offered by the aConCorde concordance program (Roberts *et al.*, 2006). A concordance program helps the user to explore vocabulary use in a Corpus by visualizing a key word in context; aConCorde goes further by allowing users to visualize clusters of words which behave significantly similarly. A third example of unsupervised learning is our entrant to the PASCAL MorphoChallenge2005 contest, the Combinatory Hybrid Elementary Analysis of Text system (Atwell and Roberts, 2006). This learns from the results of several rival systems, and chooses the most significant analysis of each input from the rival analyses.

The use of the term "significance" is not what statisticians may expect: linguistic significance is in principle determined not by statistical metrics, but in terms of "usefulness" or "naturalness" to a linguist user of the system. However, we use statistical significance metrics to approximately predict linguistic significance. Our research so far has focused on Corpus texts from a wide range of genres and languages. We hope that the LASR workshop provides an opportunity to explore applications of unsupervised learning of linguistic significance to bioinformatics datasets and problems.

References

Atwell, E. and Roberts, A. (2006). Combinatory Hybrid Elementary Analysis of Text. Submitted to MorphoChallenge2005 workshop.

PASCAL: Pattern Analysis, Statistical and Computational modelling and Learning network. (2005). Unsupervised Segmentation of Words into Morphemes Challenge. FAQ at `http://www.cis.hut.fi/morphochallenge2005/faq.shtml`.

Abu Shawar, B. and Atwell, E. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics,* **10**, 489-516.

Roberts, A., Al-Sulaiti, L. and Atwell, E. (2006). aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic. Submitted to Corpora journal.