Paradigm Relative Entropy and Discriminative Learning

Vito Pirrelli, Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo
Institute for Computational Linguistics, CNR Pisa, Italy

Introduction

The interactive role of intra-paradigmatic and inter-paradigmatic distributions has been investigated in accounting for differential effects on visual lexical recognition for both inflected (Milin et al., 2009a, 2009b) and derived words (see Kuperman et al., 2010; Bertram et al., 2005; Schreuder et al. 2003 among others). In particular, Milin and colleagues focus on the divergence between the distribution of inflectional endings within a single paradigm (measured as the entropy of the distribution of paradigmatically-related forms, or Paradigm Entropy), and the distribution of the same endings within their broader inflectional class (measured as the entropy of the distribution of inflectional endings across all paradigms, or Inflectional Entropy). They conclude that both entropic scores facilitate visual lexical recognition, but if the two distributions differ, a conflict arises, resulting in slower word recognition. Similar results are reported by Kuperman and colleagues (2010) on reading times for Dutch derived words, and are interpreted as reflecting an information imbalance between the family of the base word (e.g. *plaats* in *plaatsing*) and the family of the suffix (-*ing*).

The difference between Paradigm Entropy and Inflectional Entropy can be expressed in terms of Relative Entropy, or Kullback-Leibler divergence (D_{KL}, Kullback 1987), as follows:

1) $D_{KL}(p(e\,|s)||p(e)) = \sum_e p(e \mid s) log \frac{p(e|s)}{p(e)}$,

where $p(e \mid s)$ represents the probability of having a specific inflected form (an ending e) given a stem s, and $p(e)$ the probability of encountering e. For any specific paradigm being selected, the larger D_{KL}, the more difficult is, on average, the visual recognition of members of that paradigm.

Although these effects are clear in broad outline, no computational models of lexical processing we know of have been able to simulate them and bring them down to some underlying mechanisms of *discriminative learning* (Rescorla & Wagner 1972, Ramscar & Yarlett 2007, Baayen et al. 2011, Blevins 2016). In the present contribution, we show that principles of discriminative learning of symbolic time series go a long way in accounting for these effects, thus making an important contribution to our understanding of the human lexical processor and its sensitivity to word distributions both within and across paradigms.

Background

In Temporal Self-Organising Maps (or TSOMs: Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015), a family of neural networks based on Kohonen SOMs (Kohonen

2001), weights on a layer of temporal inter-node connections encode how strongly the currently most highly activated node or Best Matching Unit at time *t* (BMU(*t*)) is predicted by the BMU(*t-1*) at the previous time tick. A weight close to 0 on the connection between BMU(*t-1*) and BMU(*t*) indicates that the activation of BMU(*t*) is unexpected and thus somewhat surprising, given BMU(*t-1*). A weight close to 1 means that the activation is highly expected, and thus poorly informative. In TSOMs, connection weights are tuned as the result of training the map on input data, according to principles of correlative learning that are strongly reminiscent of Rescorla & Wagner (1972) discriminative equations. Given the input bigram ‘AX’, for example,

(i) the connection between BMU(‘A’) at time *t-1* and BMU(‘X’) at time *t* is strengthened (entrenchment);

(ii) the connections to BMU(‘X’) from all the other nodes are weakened (competition).

The interaction between entrenchment and competition accounts for effects of context-sensitive specialisation of map nodes for input strings. If the bigram ‘AX’ is repeatedly input to a TSOM, the map tends to develop a specialised BMU(‘X’) for ‘X’ in ‘AX’ and a highly-weighted outward connection from BMU(‘A’) to BMU(‘X’). Since node specialisation propagates through time, a TSOM is thus biased in favour of memorising input strings through BMUs structured in a word-tree, as opposed to a word-graph (Figure 1).

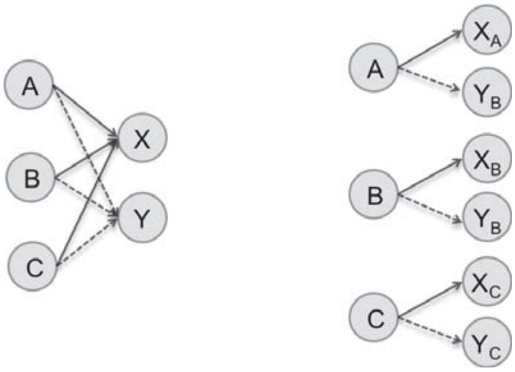

Figure 1: A TSOM trained on the three mini-paradigms ‘AX’, ‘AY’, ‘BX’, ‘BY’, ‘CX’, ‘CY’ will tend to progressively move away from a graph-like allocation of nodes to symbols (left panel) towards a tree-like allocation (right panel). The extent to which context-sensitive specialisation takes place is a function of intra-paradigmatic and inter-paradigmatic word distributions (see main text for details).

Relative entropy and paradigm learning: an experiment on mini-paradigms

The relatively simple dynamic expressed by the two learning rules (i, ii) accounts for facilitatory effects of paradigm entropy and inflectional entropy on word learning.

To illustrate, we trained a TSOM on three mini-paradigms, whose forms are obtained by combining three stems (‘A’, ‘B’ and ‘C’) with two endings (symbols ‘X’ and ‘Y’). Mini-paradigms were administered to the map on six training regimes (R1-R6, see Table 1), whose distribution was intended to control the comparative probability distribution of ‘X’ and ‘Y’, and the comparative probability distribution of the stems ‘A’, ‘B’ and ‘C’ relative to each ending. Across regimes 1-3, we kept the frequency distribution of X constant (but let it vary across paradigms), while increasing the distribution of Y both within each paradigm (R2), and across paradigms (R3). Across regimes 4-5, the frequency of Y was held constant, while X frequencies were made vary. Finally in R6 all word frequencies were put to 100. Note that, in R3 and R6, *p*(*e* | *s*) = p(*e*), i.e. the distribution of each inflected form within a paradigm equals the distribution of its ending (given its inflection class).

Results of the different training regimes are shown in Figure 2, where we plotted weights on the connection between stems ('A', 'B' and 'C') and endings ('X' and 'Y') by learning epochs, averaged over 100 repetitions of the same experiment on each regime. Results were analysed with linear mixed-effects models, with stem-ending connection weights as our dependent variable and the following three fixed effects: 1) the word probability *p*(*s*, *e*), expressed as a stem-ending combination; 2) the probability p(*e* | *s*) of a stem selecting a specific ending (or intra-paradigmatic competition), and 3) the conditional probability *p*(*s* | *e*) of a given ending being selected by a specific stem (inter-paradigmatic competition). Experiment repetitions were used as random effects. Here, we shortly summarise the main results observed.

		training regimes					
paradigm id	items	R1	R2	R3	R4	R5	R6
A	#,A,X,$	5	5	5	5	5	100
A	#,A,Y,$	5	50	50	333	333	100
B	#,B,X,$	10	10	10	10	100	100
B	#,B,Y,$	10	100	100	333	333	100
C	#,C,X,$	85	85	85	85	850	100
C	#,C,Y,$	10	100	850	333	333	100

Table 1: Frequency distribution of mini-paradigms for 6 training regimes.

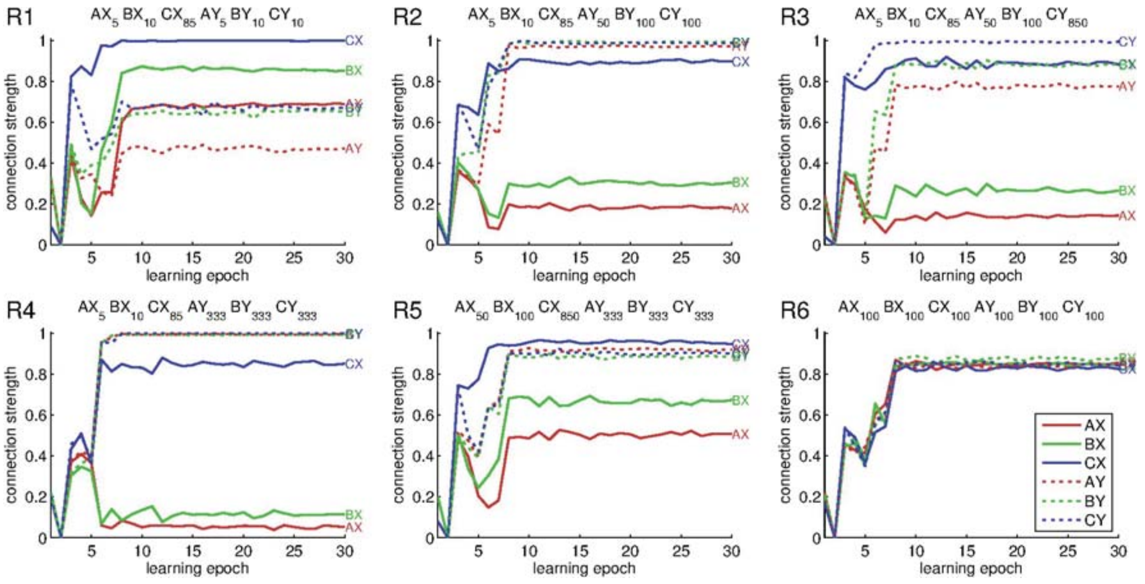

Figure 2: Developmental trends of connection strength at the stem-ending boundary under different training regimes with three mini-paradigms (R1-R6, see Table 1). Weights are plotted against the first 30 learning epochs.

Due to entrenchment (rule i), the strength of each connection at the morpheme boundary tends to be a direct function of the probability of each word form, or *p*(*s*,*e*) (see panel R3). However, other distributions interact with word frequency: connection strengths are affected by the probability of each ending *p*(*e*), with low-frequency words that contain high-frequency endings showing a stronger boundary connection than low-frequency words that contain less frequent endings (panel R1). This boosting effect is modulated by two further interactions: the conditional probability distribution *p*(*e* | *s*), with connections to 'X' suffering from an increase in the probability mass of 'Y' (panels R2 and R4), and the competition between words selecting the same ending (rule ii), modulated by the entropy of the conditional probability distribution *p*(*s* | *e*), or *H*(*s* | *e*) (panels R4 and R5). In particular, if we control *H*(*s*), i.e. the distribution of paradigms in the input data, the entropy *H*(*s* | *e*) is expressed analytically by the following equation:

$$2)\quad H(s \mid e) = H(s) - \sum_{s,e} p(s,e) log \frac{p(s,e)}{p(s)p(e)} \,,$$

where $\sum_{s,e} p(s,e) log \frac{p(s,e)}{p(s)p(e)}$ is known as Mutual Information. Using the Bayesian equality *p*(s,e) = *p*(s)*p*(e|s), we can rewrite equation (2) above as follows:

$$3)\quad H(s \mid e) = H(s) - \sum_{s} p(s) \sum_{e} p(e \mid s) log \frac{p(e|s)}{p(e)} \,,$$

where $\sum_{e} p(e \mid s) log \frac{p(e|s)}{p(e)}$ is the Kullback-Leibler divergence $D_{KL}(p(e|s)||p(e))$ between *p*(*e* | *s*) and *p*(e) (Eq. 1 above). Equation (3) shows that *H*(*s* | *e*) is maximised by minimising the average divergence between the intra-paradigmatic distribution *p*(*e* | *s*) of the endings given a stem, and the marginal distribution *p*(*e*) of the endings. In other words, verb paradigms are learned more accurately by a TSOM when, on average, the distribution *p*(*e* | *s*) of the forms within each paradigm approximates the marginal distribution of each ending in the corresponding conjugation class (compare R4 and R6). This behaviour, accounted for by the interaction of entrenchment and competition in discriminative learning, is in line with the facilitation effects reported for visual lexical recognition of inflected words and reading times of derived words. Besides, the evidence is compatible with more extensive experiments on German and Italian verbs (Marzi et al. 2014), showing that, for comparable cumulative frequencies, uniform distributions in training data (R6) facilitate paradigm acquisition.

References

Baayen, R.H., Milin, P., Đurđević, D.F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3), 438.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000c). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 489–511.

Blevins, J.P. (2016). Word and paradigm morphology. Oxford University Press.

Colé, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, 28, 1–13.

Ferro M., Marzi, C., & Pirrelli, V. (2011). A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning*. Lingue e Linguaggio*, X(2), 209-226.

Kohonen, T. (2001). *Self-Organizing Maps*. Heidelberg, Springer-Verlag.

Kullback, S. (1987). Letter to the editor: The Kullback-Leibler distance. The American Statistician, 41(4), 340-341.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, 62(2), 83-97.

Marzi, C., Ferro, M., & Pirrelli, V. (2014). Morphological structure through lexical parsability. Lingue e Linguaggio, XIII(2), 263-290.

Milin, P., Đurđević, D.F., & del Prado Martín, F.M. (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. Journal of Memory and Language, 60(1), 50-64.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R.H. (2009b). Words and paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J.P. Blevins & Blevins, J. (Eds.) Analogy in grammar: Form and acquisition, 214-252. Oxford University Press.

Pirrelli, V., Ferro, M., & Marzi, C. (2015). Computational complexity of abstractive morphology. In Baerman, M., Brown, D. & Corbett, G. (Eds.), Understanding and Measuring Morphological Complexity. Oxford: Oxford University Press. 141-166.

Ramscar, M., & Yarlett, D, (2007) Linguistic Self-Correction in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition. Cognitive Science, 31, 927-960.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory, 2, 64-99.