Journal of King Saud University – Computer and Information Sciences 29 (2017) 179–188

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Arabic word processing and morphology induction through adaptive memory self-organisation strategies

Claudia Marzi *, Marcello Ferro, Ouafae Nahli

Institute for Computational Linguistics, National Research Council (ILC-CNR), Pisa, Italy

ARTICLE INFO

Article history:
Received 29 January 2016
Revised 30 June 2016
Accepted 13 November 2016
Available online 19 November 2016

Keywords:
Non-concatenative morphological structure
Word recoding and processing
Lexical storage and access
Topological alignment
Synchronisation
Self-Organising Maps

ABSTRACT

Aim of the present study is to model the human mental lexicon, by focussing on storage and processing dynamics, as lexical organisation relies on the process of input recoding and adaptive strategies for long-term memory organisation. A fundamental issue in word processing is represented by the emergence of the morphological organisation level in the lexicon, based on paradigmatic relations between fully-stored word forms. Morphology induction can be defined as the task of perceiving and identifying morphological formatives within morphologically complex word forms, as a function of the dynamic interaction between lexical representations and distribution and degrees of regularity in lexical data.

In the computational framework we propose here (TSOMs), based on Self-Organising Maps with Hebbian connections defined over a temporal layer, the identification/perception of surface morphological relations involves the alignment of recoded representations of morphologically-related input words. Facing a non-concatenative morphology such as the Arabic inflectional system prompts a reappraisal of morphology induction through adaptive organisation strategies, which affect both lexical representations and long-term storage.

We will show how a strongly adaptive self-organisation during training is conducive to emergent relations between word forms, which are concurrently, redundantly and competitively stored in human mental lexicon, and to generalising knowledge of stored words to unknown forms.

1. Introduction

One of the fundamental issues in defining word storage and processing is modelling the emergence of the morphological organisation level in the human lexicon, based on paradigmatic relations between fully-stored word forms.

The task of inducing morphological knowledge from lexical data can be defined as the task of singling out morphological formatives from surface word forms. Operationally, the task consists of the following steps: (i) finding structure in word forms, and (ii) grouping word forms on the basis of shared structure. Originally defined by Harris (1955) as a battery of "discovery procedures" of unclassified training data on the basis of purely formal algorithms, morphology induction mirrors the interplay between structured representation and the recoding process.

In spite of their different algorithms, both supervised and unsupervised machine learning models make a priori assumptions on the nature of the task of morphology induction. Supervised algorithms tend to rely on specific assumptions on word representations. Indeed, for most European languages, we can construe a fixed-length vector representation that aligns input words to the right, since inflection in those languages typically involves suffixation and sensitivity to morpheme boundaries. However, this type of representation presupposes considerable a priori knowledge of the morphology of the target language and does not possibly work with prefixation, circumfixation and non-concatenative morphological processes in general.

On the other hand, most current unsupervised algorithms model morphology learning as a segmentation task (Hammarström and Borin, 2011), assuming a hard-wired linear correspondence between sub-lexical strings and morphological structure. Once more, non-concatenative morphologies can hardly be segmented into linearly concatenated morphemes.

* Corresponding author at: via G. Moruzzi 1, 56124 Pisa, Italy.
E-mail addresses: claudia.marzi@ilc.cnr.it (C. Marzi), marcello.ferro@ilc.cnr.it (M. Ferro), ouafae.nahli@ilc.cnr.it (O. Nahli).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.11.006
1319-1578/© 2016 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

In line with recent psycholinguistic evidence on peripheral levels of automatic morphology segmentations (Crepaldi et al., 2010; Rastle and Davis, 2008; Velan and Frost, 2011), modelling human lexical processing and storage should rely on algorithms more valued for their general capacity to adapt themselves to the morphological structure of a target language, rather than for the strength of their inductive morphological bias.

We show that the same morphology induction algorithm, with an identical setting of initial parameters and a comparable set of assumptions concerning input representations, is able to successfully deal with as diverse inflectional systems as, for example, Italian, German and Arabic, and with diverse morphological phenomena within the same language (e.g. suffixation, prefixation, infixation and combination thereof in the Arabic verbal inflection). We suggest that a principled approach to these issues should be able to replicate some fundamental abilities lying at the heart of the human language processor: (i) recode and maintain time series of symbolic units (e.g. letters, phonological symbols, morphemes, or words) in the so-called working memory, (ii) transfer and organise these representations in the long-term memory, (iii) map input representations onto lexical representations for access and recall them in language usage, (iv) generalise knowledge of stored words to unknown forms.

Firstly, we outline the theoretical background for the present work (Section 2), the computational architecture (Section 3) adopted for our experiments, together with the analysis techniques implemented to inspect the emergence of morphological structure. Materials, methods, and results are then illustrated and analysed (Section 4), focussing on how a strongly paradigmatic co-organisation and co-activation facilitate morphological learning, extension and generalisation. A general discussion (Section 5) follows, summarising our results in the framework of an integrative model for memory, processing and access strategies.

2. Theoretical background

2.1. Recoding and memory

A fundamental characteristic of the human language faculty is the ability to retain sequences of symbolic units in the long-term memory, to access them in recognition and production, and to find similarities and differences among them. Traditionally, lexical acquisition and processing have been modelled in terms of basic mechanisms of human memory for serial order, as proposed in the vast literature on immediate serial recall and visual word recognition (e.g. Henson (1999), Davis (2010); for detailed reviews). Some of the earliest psychological accounts of serial order assume that item sequences are represented as temporal chains made up of stimulus–response links. However, it can be difficult to temporally align word forms of differing lengths, thus preventing recognition of shared sequences between morphologically-related forms (Davis and Bowers, 2004), in particular in case of abstract bound morpheme like the discontinuous symbols of consonantal root in Arabic language (Boudelaa and Marslen-Wilson, 2004). Conventionally, the task of identifying morphological formatives within morphologically complex word forms has been taken to model morphology induction. Accordingly, there is a general problem that any such model has to address and that appears to be crucial for morphology induction: the word alignment issue. The problem arises whenever familiar patterns are presented in novel arrangements, as when speakers of English are able to recognise the word *book* in *handbook*, or Arabic speakers can track down the verb root *k-t-b* in *kataba* ('he wrote') and *yaktubu* ('he writes'). No position-specific letter coding scheme can account for such ability.

In Davis' spatial encoding (2010), a letter in a string is represented as a two-dimensional signal. The identity of the letter is described as a Gaussian activity function whose maximum value is centred on the letter's actual position and decreases continuously as we move away from that position either rightwards or leftwards. The function defines a confidence level on the position of the letter in question. String matching is continuously weighted by levels of positional confidence, thus enforcing a form of fuzzy matching. However, the approach, as most other psycho-cognitively inspired models such as the "open-bigram coding" model (Grainger and van Heuven, 2003), the "start–end" model (Henson, 1998) and the "primacy model" (Page and Norris, 1998) among others, is chiefly recognition-oriented and is not readily amenable to model human word processing, morphology induction and generalisation.

2.2. Paradigmatic relations

One of the most prominent issues in modelling word acquisition and processing is represented by the emergence of a level of morphological organisation in the human lexicon. In the perspective of adaptive strategies for lexical acquisition and processing based on emergent morphological relations between fully-stored word forms (defined as an *abstractive* approach after Blevins, 2006), paradigmatic[1] relations can be accounted for as the result of long-term entrenchment of neural circuits (chains of time-stamped memory nodes) that are repeatedly being activated.

Discontinuous morphological formatives – e.g. roots in the Arabic inflectional system – or discontinuous morphological processes – e.g. circumfixation in German past participles, Arabic imperfective forms – represent a challenge to the notion that identical structures are responded to by topologically adjacent nodes. The root *k-t-b* is, for example, dramatically misaligned in *kataba* and *yaktubu*, and this may keep the nodes responding to the root in two – or more – words far apart on the map. Likewise, *machen* ('make, we/they make') and *gemacht* ('made' past participle) are temporally misaligned although sharing the same stem.

In previous works (Marzi et al., 2012c, 2014), we analysed the paradigmatic organisation of the inflectional morphology of German and Italian, by focussing on how different types of related intra- and inter-paradigmatic families induce a strongly paradigm-related co-organisation and co-activation so as to facilitate paradigmatic extension and generalisation. In the framework of Temporal Self-Organising Maps (TSOMs), a variant of classical SOMs (Kohonen, 2001) augmented with re-entrant Hebbian connections defined over a temporal layer, which can encode probabilistic expectations upon incoming stimuli (Koutnik, 2007; Ferro et al., 2010, 2011; Pirrelli et al., 2011; Marzi et al., 2012a,b), we showed how deeply entrenched chains of nodes are concurrently activated by morphologically related word forms. In particular, we highlighted how, from a lexical standpoint, TSOMs exhibit a straightforward correlation between morphological segmentation and topological organisation of memory nodes.

3. The computational framework

TSOMs are two-dimensional grids of artificial memory nodes, which are not wired-into maximally respond to specific symbols

[1] A verb paradigm represents a family of inflected variants of the same lexical exponent (e.g. *play, plays, paying, played*), whereas inflectional classes denote families of similarly inflected forms (e.g. *played*, *walked, arrived*). The role of paradigmatic relations is considered, in the theoretical and psycho-cognitive literature, as central in organisation of word forms in speakers' mental lexicon, facilitating lexical access and storage (Bybee and Slobin, 1982; Bybee and Moder, 1983; Baayen et al., 1997; among others).