How "deep" is learning word inflection?

Franco Alberto Cardillo
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
francoalberto.cardillo@ilc.cnr.it

Marcello Ferro
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
marcello.ferro@ilc.cnr.it

Claudia Marzi
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
claudia.marzi@ilc.cnr.it

Vito Pirrelli
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
vito.pirrelli@ilc.cnr.it

Abstract

English. Machine learning offers two basic strategies for morphology induction: lexical segmentation and surface word relation. The first one assumes that words can be segmented into morphemes. Inducing a novel inflected form requires identification of morphemic constituents and a strategy for their recombination. The second approach dispenses with segmentation: lexical representations form part of a network of associatively related inflected forms. Production of a novel form consists in filling in one empty node in the network. Here, we present the results of a recurrent LSTM network that learns to fill in paradigm cells of incomplete verb paradigms. Although the process is not based on morpheme segmentation, the model shows sensitivity to stem selection and stem-ending boundaries.

Italiano. *La letteratura offre due strategie di base per l'induzione morfologica. La prima presuppone la segmentazione delle forme lessicali in morfemi e genera parole nuove ricombinando morfemi conosciuti; la seconda si basa sulle relazioni di una forma con le altre forme del suo paradigma, e genera una parola sconosciuta riempiendo una cella vuota del paradigma. In questo articolo, presentiamo i risultati di una rete LSTM ricorrente, capace di imparare a generare nuove forme verbali a partire da forme gïï¿œ note non segmentate. Ciononostante, la rete acquisisce una conoscenza implicita del tema verbale e del confine con la terminazione flessionale.*

1 Introduction

Morphological induction can be defined as the task of singling out morphological formatives from fully inflected word forms. These formatives are understood to be part of the morphological lexicon, where they are accessed and retrieved, to be recombined and spelled out in word production. The view requires that a word form be segmented into meaningful morphemes, each contributing a separable piece of morpho-lexical content. Typically, this holds for regularly inflected forms, as with Italian *cred-ut-o* 'believed' (past participle, from CREDERE), where *cred-* conveys the lexical meaning, and *-ut-o* is associated with morpho-syntactic features. A further assumption is that there always exists an underlying *base* form upon which all other forms are spelled out. In an irregular verb form like Italian *appes-o* 'hung' (from APPENDERE), however, it soon becomes difficult to separate morpholexical information (the verb stem) from morpho-syntactic information.

A different formulation of the same task assumes that the lexicon consists of fully-inflected word forms and that morphology induction is the result of finding out implicative relations between them. Unknown forms are generated by redundant analogy-based patterns between known forms, along the lines of an analogical proportion such as: *rendere* 'make' :: *reso* 'made' = *appendere* 'hang' :: *appeso* 'hung'. Support to this view comes from developmental psychology, where words are understood as the foundational elements of language acquisition, from which early grammar rules emerge epiphenomally (Tomasello, 2000; Goldberg, 2003). After all, children appear to be extremely sensitive to subregularities holding between inflectionally-related forms (Bittner et al., 2003; Colombo et al., 2004;

Dąbrowska, 2004; Orsolini and Marslen-Wilson, 1997; Orsolini et al., 1998). Further support is lent by neurobiologically inspired computer models of language, blurring the traditional dichotomy between processing and storage (Elman, 2009; Marzi et al., 2016). In particular we will consider here the consequences of this view on issues of word inflection by recurrent Long Short Term Memory (LSTM) networks (Malouf, in press).

2 The cell-filling problem

To understand how word inflection can be conceptualised as a word relation task, it is useful to think of this task as a *cell-filling problem* (Blevins et al., 2017; Ackerman and Malouf, 2013; Ackerman et al., 2009). Inflected forms are traditionally arranged in so-called *paradigms*. The full paradigm of CREDERE 'believe' is a labelled set of all its inflected forms: *credere, credendo, creduto, credo* etc. In most cases, these forms take one and only one *cell*, defined as a specific combination of tense, mood, person and number features: e.g. *crede*, PRES IND, 3S. In all languages, words happen to follow a Zipfian distribution, with very few high-frequency words, and a vast majority of exceedingly rare words (Blevins et al., 2017). As a result, even high-frequency paradigms happen to be attested partially, and learners must then be able to generalise incomplete paradigmatic knowledge. This is the cell-filling problem: given a set of attested forms in a paradigm, the learner has to guess other missing forms in the same paradigm.

The task can be simulated by training a learning model on a number of partial paradigms, to then complete them by generating missing forms. Training consists of <lemma_paradigm cell, inflected form> pairs. A lemma is not a form (e.g. *credere*), but a symbolic proxy of its lexical content (e.g. CREDERE). Word inflection consists of producing a fully inflected form given a known lemma and an empty paradigm cell.

2.1 Methods and materials

Following Malouf (in press), our LSTM network (Figure 1) is designed to take as input a lemma (e.g. CREDERE), a set of morphosyntactic features (e.g. PRES_IND, 3, S) and a sequence of symbols (<*crede*>)[1] one symbol s_t at a time, to output a probability distribution

[1] '<' and '>' are respectively the start-of-word and the end-of-word symbols

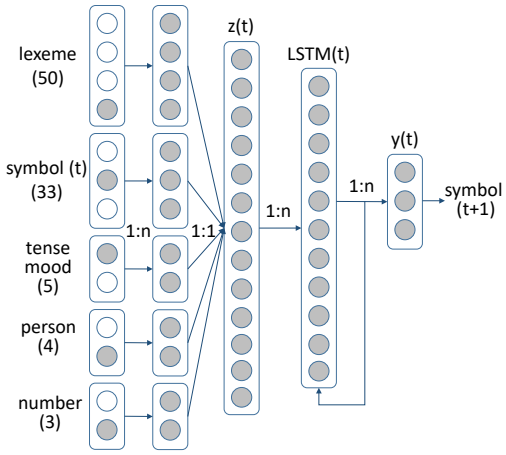

Figure 1: The network architecture. The input vector dimension is shown in brackets. Trainable dense projection matrices are shown as $1:n$, and concatenation as $1:1$.

over the upcoming symbol s_{t+1} in the sequence: $p(s_{t+1}|s_t$,CREDERE, PRES_IND, 3, S). To produce the form <*crede*>, we take the start symbol '<' as s_1, use s_1 to predict s_2, then use the predicted symbol to predict s_3 and so on, until '>' is predicted. Input symbols are encoded as mutually orthogonal one-hot vectors with as many dimensions as the overall number of different symbols used to encode all inflected forms. The morphosyntactic features of tense, person and number are given different one-hot vectors, whose dimensions equal the number of different values each feature can take.[2] All input vectors are encoded by trainable dense matrices whose outputs are concatenated into the projection layer $z(t)$, which is in turn input to a layer of LSTM blocks (Figure 1). The layer takes as input both the information of $z(t)$, and its own output at *t–1*. Recurrent LSTM blocks are known to be able to capture long-distance relations in time series of symbols (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997; Jozefowicz et al., 2015), avoiding classical problems with training gradients of Simple Recurrent Networks (Jordan, 1986; Elman, 1990).

We tested our model on two comparable sets of Italian and German inflected verb forms (Table 1), where paradigms are selected by sampling the highest-frequency fifty paradigms in two reference corpora (Baayen et al., 1995; Lyding et al., 2014). For both languages, a fixed set of cells was

[2] Note that an extra dimension is added when a feature can be left uninstatiated in particular forms, as is the case with person and number features in the infinitive.

language	alphabet size	max len	cells	reg / irreg paradigms	forms
German	27	13	15	16 / 34	750
Italian	21	14	15	23 / 27	750

Table 1: The German and Italian datasets.

chosen from each paradigm: all present indicative forms (n=6), all past tense forms (n=6), infinitive (n=1), past participle (n=1), German present participle/Italian gerund (n=1).[3] The two sets are inflectionally complex: they exhibit extensive stem allomorphy and a rich set of affixations, including circumfixation (German *ge-mach-t* 'made', past participle). Most importantly, the distribution of stem allomorphs is entirely accountable in terms of equivalence classes of cells, forming morphologically heterogenous, phonologically poorly predictable, but fairly stable sub-paradigms (Pirrelli, 2000). Selection of the contextually appropriate stem allomorph for a given cell thus requires knowledge of the form of the allomorph and of its distribution within the paradigm.

3 Results and discussion

To meaningfully assess the relative computational difficulty of the cell-filling task, we calculated a simple baseline performance, with 695 forms of our original datasets selected for training, and 55 for testing.[4]For this purpose, we used the baseline system for Task 1 of the CoNLL-SIGMORPHON-2017 Universal Morphological Reinflection shared task.[5]The model changes the infinitive into its inflected forms through rewrite rules of increasing specificity: *e.g.* two Italian forms such as *badare* 'to look after' and *bado* 'I look after' stand in a BASE :: PRES_IND_3S relation. The most general rule changing the former into the latter is *-are -> -o*, but more specific rewrite rules can be extracted from the same pair:

[3]The full data set is available at `http://www.comphyslab.it/redirect/?id=clic2017_data`. Each training form is administered once per epoch, and the number of epochs is a function of a "patience" threshold. Although a uniform distribution is admittedly not realistic, it increases the entropy of the cell-filling problem, to define some sort of upper bound on the complexity of the task.

[4]Test forms were selected to constitute a benchmark for evaluation. We made it sure that a representative sample of German and Italian irregulars were included for evaluation, provided that they could be generalised on the basis of the training data available.

[5]`https://github.com/sigmorphon/conll2017`(written by Mans Hulden).

German test	all	regs	irregs
CoNLL baseline	0.4	0.81	0.23
128-blocks	0.68	0.79	0.64
256-blocks	**0.75**	**0.89**	**0.69**
512-blocks	0.71	0.84	0.66

Italian test	all	regs	irregs
baseline	0.65	0.9	0.5
128-blocks	0.61	0.84	0.47
256-blocks	0.63	0.83	0.51
512-blocks	**0.69**	**0.92**	**0.54**

Table 2: Per-word accuracy in German and Italian. Overall scores for the three word classes are averaged across 10 repetitions of each LSTM type.

-dare -> -do, *-adare -> -ado*, *-badare -> -bado*. The algorithm then generates the PRES_IND_3S of - say - *diradare* 'thin out', by using the rewrite rule with the longest left-hand side matching *diradare* (namely *-adare* > *-ado*). If there is no matching rule, the base is used as a default output.

The algorithm proves to be effective for regular forms in both languages (Table 2). However, per-word accuracy drops dramatically on German irregulars (0.23), and Italian irregulars (0.5). The same table shows accuracy scores on test data obtained by running 128, 256 and 512 LSTM blocks. Each model instance was run 10 times, and overall per-word scores are averaged across repetitions.[6]

The CoNLL baseline is reminiscent of Albright and Hayes' (2003) Minimal Generalization Learner, inferring Italian infinitives from first singular present indicative forms (Albright, 2002). In the present case, however, the inference goes from the infinitive (base) to other paradigm cells. The inference is much weaker in German, where stem allomorphy is more consistently distributed within each paradigm. In Appendix, Table 3 contains a list of all German forms wrongly produced by the CoNLL baseline, together with per-word accuracy of our models. Most wrong forms are inflected forms requiring ablaut, which turn out to be over-regularised by the CoNLL baseline (e.g. **stehtet* for *standet*, **beginntet* for *begannt*). It appears that, in German, a purely syntagmatic approach to word production, deriving all inflected forms from an underlying base, has a strong bias towards over-regularisation. Simply put, the orthotactic/phonotactic structure of the German stem

[6]The per-word score is 1 (correct), or 0 (wrong).

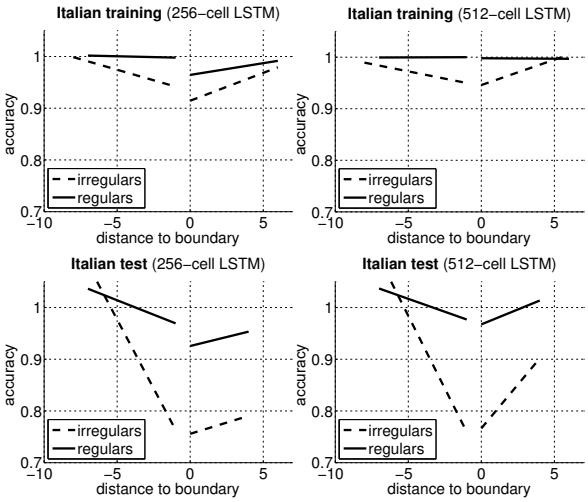

Figure 2: Marginal plots of the interaction between distance to morpheme boundary, stem/inflectional ending, inflectional regularity, stem length and suffix length (fixed effects) in a LME model fitting per-symbol accuracy by a 256-block (left) and a 512-block (right) RNN on training (top) and test (bottom) Italian data. Random effects are model repetitions and word forms.

is less criterial for stem allomorphy than the Italian one. LSTMs are considerably more robust in this respect. Memory resources allowing, they can keep track of local syntagmatic constraints as well as more global, paradigmatic constraints, whereby *all* paradigmatically-related forms contribute to fill in gaps in the same paradigm. For example, knowledge that a paradigm contains a few stem allomorphs is good reason for an LSTM to produce a stem allomorph in other (empty) cells. The more systematic the distribution of stem alternants is across the paradigm, the easier for the learner to fill in empty cells. German conjugation proves to be paradigmatically well-behaved.

An LSTM recurrent network has no information about the morphological structure of input forms. Due to the predictive nature of the production task and the LSTM re-entrant layer, however, the network develops a left-to-right sensitivity to upcoming symbols, with per-symbol accuracy being a function of the network confidence about the next output symbol. To assess the correlation between per-symbol accuracy and "perception" of the morphological structure, we used a Linear Mixed Effects (LME) model of how well structural features of German and Italian verb forms interpolate the "average" network accuracy in producing an upcoming symbol (1 for a hit, 0 for a miss) in both training and test. The marginal plots of Figure 2 show that there is a clear structural effect of the distance to the stem-ending boundary of the symbol currently being produced, over and above the length of the input string. Besides, stems and suffixes of regulars exhibit different accuracy slopes compared with stems and suffixes of irregulars. Intuitively, production of an inflected form by a LSTM network is fairly easy at the beginning of the stem, but it soon gets more difficult when approaching the morpheme boundary, particularly with irregulars. Accuracy reaches the minimum value on the first symbol of the inflectional ending, which marks a point of structural discontinuity in an inflected verb form. From that position, accuracy starts increasing again, showing a characteristically V-shaped trend. Clearly, this trend is more apparent with test words (Figure 2, bottom), where stems and endings are recombined in novel ways. The same results hold for German. On the other hand, no evidence of structure sensitivity was found in a LME model of the baseline output for both German and Italian.

The cell-filling problem is an ecological, developmentally motivated task, based on evidence of fully inflected forms. Although other (simpler) models have been proposed to account for form-meaning mapping in Morphology (Baayen et al., 2011; Plaut and Gonnerman, 2000, among others), we do not know of any other artificial neural networks that can simulate word inflection as a cell-filling task. Unlike more traditional connectionist architectures (Rumelhart and McClelland, 1986), recurrent LSTMs do not presuppose the existence of underlying base forms, but they learn possibly alternating stems upon exposure to full forms. Admittedly, the use of orthogonal one-hot vectors for lemmas, unigram temporal series for inflected forms, and abstract morpho-syntactic features as a proxy of context-sensitive functional agreement effects, are crude representational short-hands. Nonetheless, in tackling the task, LSTMs prove to be able to orchestrate "deep" knowledge about word structure, well beyond pure surface word relations: namely stem-affix boundaries, paradigm organisation and degrees of regularity in stem formation. Acquisition of different inflectional systems may require a different balance of all these pieces of knowledge.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Adam Albright. 2002. Islands of reliability for regular morphology: Evidence from italian. *Language*, pages 684–709.

Harald R. Baayen, P. Piepenbrock, and L. Gulikers, 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Harald R. Baayen, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438–481.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Dagmar Bittner, Wolfgang U. Dressler, and Marianne Kilani-Schoch, editors. 2003. *Development of Verb Inflection in First Language Acquisition: a cross-linguistic perspective*. Mouton de Gruyter, Berlin.

James P. Blevins, Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins, and Huba Bartos, editors, *Morphological Paradigms and Functions*. Brill, Leiden.

Lucia Colombo, Alessandro Laudanna, Maria De Martino, and Cristina Brivio. 2004. Regularity and/or consistency in the production of the past participle? *Brain and language*, 90(1):128–142.

Ewa Dąbrowska. 2004. Rules or schemas? evidence from polish. *Language and cognitive processes*, 19(2):225–271.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Adele E Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Michael Jordan. 1986. Serial order: A parallel distributed processing approach. Technical Report 8604, University of California.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell' Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisá corpus of italian web texts. Proceedings of the 9th Web as Corpus Workshop (WaC-9)@ EACL 2014, pages 36–43. Association for Computational Linguistics.

Robert Malouf. in press. Generating morphological paradigms with a recurrent neural network. *Morphology*.

Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo, and Vito Pirrelli. 2016. Effects of frequency and regularity in an integrative model of word storage and processing. *Italian Journal of Linguistics*, 28(1):79–114.

Margherita Orsolini and William Marslen-Wilson. 1997. Universals in morphological representation: Evidence from italian. *Language and Cognitive Processes*, 12(1):1–47.

Margherita Orsolini, Rachele Fanari, and Hugo Bowles. 1998. Acquiring regular and irregular inflection in a language with verb classes. *Language and cognitive processes*, 13(4):425–464.

Vito Pirrelli. 2000. *Paradigmi in morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano*. Istituti Editoriali e Poligrafici Internazionali, Pisa.

David C Plaut and Laura M Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4/5):445–485.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of english verbs. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing. Explorations in the Microstructures of Cognition*, volume 2 Psychological and Biological Models, pages 216–271. MIT Press.

Michael Tomasello. 2000. The item-based nature of children's early syntactic development. *Trends in cognitive sciences*, 4(4):156–163.

Appendix A. Comparative test results

base form	target	CoNNL **baseline**	LSTM 128	LSTM 256	LSTM 512
bleiben	bliebt	**blieb**	0	0	0
dï¿œrfen	gedurft	**gedï¿œrfen**	0.2	0.2	0
sein	seiend	**seind**	0	0	0
mï¿œssen	gemusst	**gemï¿œssen**	0.2	0.1	0.1
bestehen	bestandet	**bestehtet**	0.5	0.6	0.2
sprechen	spricht	**sprecht**	0	0.5	0.2
geben	gibt	**gebt**	0	0.3	0.3
sehen	siehst	**sehst**	0	0	0.3
tun	tatet	**tut**	0.3	0	0.3
stehen	standet	**stehtet**	0.2	0.1	0.4
fahren	fï¿œhrst	**fahrst**	0.2	0.6	0.5
finden	fandet	**findet**	0.8	0.6	0.6
dï¿œrfen	darf	**dï¿œrfe**	0.5	0.7	0.7
fahren	fuhrst	**fahrtest**	0.4	0.4	0.7
beginnen	begannt	**beginntet**	0.6	0.9	0.8
kommen	kamst	**kommst**	1	1	0.8
liegen	lagt	**liechtet**	0.5	0.9	0.8
sehen	saht	**sehtet**	0.9	0.8	0.8
bringen	brachtet	**brinchtet**	1	1	0.9
fragen	fragtet	**frugt**	1	1	0.9
gehen	gingt	**gehtet**	0.9	1	0.9
haben	hattet	**habt**	1	1	0.9
nehmen	nahmt	**neht**	0.9	1	0.9
nennen	nanntet	**nenntet**	0.8	1	0.9
sagen	sagtet	**sugt**	1	1	0.9
tragen	trï¿œgst	**tragst**	0.9	0.9	0.9
bitten	baten	**bitten**	1	1	1
denken	dachtest	**denkest**	1	1	1
geben	gabst	**gebst**	1	1	1
scheinen	schienst	**scheintest**	0.8	1	1
setzen	setztet	**setzet**	1	1	1
sprechen	sprachst	**sprechtest**	0.8	1	1
werden	wurdet	**werdet**	0.9	1	1

Table 3: Comparative results for the 33 German verb forms that are wrongly inflected by the CoNNL baseline (highlighted in bold). In most cases, forms are over-regularised. Results are ordered by increasing accuracy of the 512-block LSTM model. Accuracy scores are given per word, and averaged across repetitions of each LSTM model in the [0, 1] range: '0' means that the output is wrong in all model repetitions, '1' that it is always correct. The most accurate results are provided by the 256-block LSTM model.