Approcci neurali alla flessione

Relazione di Psicolinguistica computazionale

Eleonora Rossi

2020/21

Indice

1 Introduzione al problema

La seguente relazione ha lo scopo di esaminare in modo critico due approcci che affrontano il problema della flessione verbale, alla luce dei requisiti di codifica di input e output e dei vincoli dettati dall'architettura della rete scelta. Tali approcci sono stati esaustivamente descritti nei rispettivi articoli:

- Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. In Rumelhart, D. E., McClelland, J.L., and the PDP Research Group (eds.), Parallel Distributed Processing. Explorations in the Microstructures of Cognition, volume 2, Psychological and Biological Models. MIT Press.
- Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. Morphology.

I due articoli appartengono a tempi storici differenti, per questo, pur affrontando entrambi il problema della flessione con modelli neurali, presentano tecniche e approcci molto diversi.

Andando nello specifico, il problema che intendono trattare consiste nel tentativo di realizzare una modello neurale capace di apprendere proprietà di una lingua morfologicamente complessa. Sia Malouf che Rumelhart e McClelland hanno una visione astrattiva [1] della competenza morfologica, ovvero il lessico è visto come sistema dinamico e aperto costituito da forme flesse piene a partire dalle quali è possibile astrarre sotto-costituenti. Si ritiene che si possa produrre/comprendere forme nuove attraverso processi di estensione analogica intra ed inter paradigmatica.

1.1 Un'alternativa al modello di *Explicit inaccessible rule*

Nel caso di Ruhmelhart e McClelland si intende proporre una valida alternativa al cosiddetto modello delle ***explicit inaccessible rule*** secondo cui le regole, necessarie ai parlanti per poter giudicare la correttezza o meno di una espressione di una lingua, sono interiorizzate in forme esplicite, ma non possono essere descritte verbalmente poiché codificate in un linguaggio speciale che solo il sistema di *processing* linguistico è in grado di comprendere.

Secondo Rumelhart & McClelland il meccanismo che processa la lingua e permette di dare giudizi di correttezza in realtà è caratterizzato da schemi sub-simbolici di regolarità. L'intenzione di questi due autori è di mostrare la correttezza della loro intuizione attraverso un semplice modello *Parallel distributed processing*. In un precedente lavoro hanno infatti dimostrato che questo tipo di modello legato all'approccio del connessionismo:

- fornisce un meccanismo sufficiente a catturare un comportamento lecito senza aver bisogno di nessuna regola esplicita ma inaccessibile
- fornisce una spiegazione alternativa alle regole esplicite ma inaccessibili per quanto riguarda la conoscenza implicita di regole

Questo lavoro tuttavia non spiega in maniera esaustiva il motivo per il quale molti dettagli del comportamento della lingua e del suo processo di acquisizione sembrano propendere verso un sistema a regole esplicite. Questo problema è stato affrontato dai due autori in questo articolo analizzando un fenomeno spesso utilizzato per dimostrare l'acquisizione di regole linguistiche ovvero il processo di acquisizione dell'uso del *Past tense* dei verbi inglesi da parte dei bambini. In particolare hanno implementato un modello PDP (*parallel distributed processing*) che apprende autonomamente a comportarsi secondo le regole, imitando le tendenze generali dei bambini nel processo di acquisizione del *past tense*.

1.2 Una soluzione al *Paradigm Cell Filling Problem*

L'articolo di Malouf invece parte da premesse piuttosto diverse. Esistono due approcci all'induzione morfologica: l'approccio costruttivo secondo cui la segmentazione delle forme lessicali in morfemi genera parole nuove ricombinando i morfemi conosciuti e un approccio astrattivo [1] il quale sostiene che, sulla base di relazioni di analogia e implicazione di una forma con le altre del suo paradigma, è possibile generare una forma nuova non conosciuta riempiendo una cella vuota del paradigma.

Andando più nello specifico, Malouf sfrutta quindi la strategia dei modelli morfologici *Words and Paradigms* secondo cui i sistemi flessivi sono insiemi di paradigmi e nuove forme possono essere create tramite analogie (o relazioni di implicazione) con forme precedentemente incontrate, per realizzare una rete neurale ricorrente (***recurrent neural network***) che apprenda paradigmi di una lingua complessa basandosi su input parziali e casuali (esattamente come farebbe un parlante). A differenza del modello di Rumelhart e McClelland la presenza di ricorrenze interne fa sì che la rete possa tener conto della natura sequenziale/temporale dei dati in input e catturare relazioni distanti, migliorando così le sue capacità predittive.

Per fare ciò l'autore intende trovare una soluzione al cosiddetto ***Paradigm Cell Filling Problem*** (Ackerman et al. 2009): dato un insieme di forme conosciute in un paradigma, il parlante deve essere in grado di riconoscere e produrre forme flesse mancanti nello stesso paradigma. Questa formalizzazione permette di comprendere come è possibile concettualizzare l'idea di flessione di parola in relazione ad altre parole dello stesso paradigma: in lingue morfologicamente complesse le forme flesse sono organizzate in paradigmi (un paradigma è costituito da tutte le forme flesse di una determinata parola), i quali sono costituiti da celle. Le celle paradigmatiche sono caratterizzate da una combinazione di caratteristiche morfosintattiche come il tempo, il modo, la persona e il numero.
Un modo per poter risolvere il problema del *Paradigm Cell Filling* è individuare una rete che, una volta appreso un paradigma parziale, è in grado di generare le forme mancanti.[1].

Dunque ricapitolando, se l'obiettivo di Rumelhart e McClelland è quello di simulare il processo di acquisizioni del paste tense di un bambino (e catturare all'evenienza altri aspetti significativi), Malouf intende invece mostrare che le RNN, in particolare le reti LSTM ricorrente, sono un ottimo strumento per analizzare morfologie flessive complesse.

2 Analisi dei due approcci

Come è possibile già notare, gli approcci si differenziano per il diverso tipo di rete neurale utilizzata: mentre Rumelhart e McClelland hanno implementato un semplice *Perceptron*, Malouf ha optato per una rete neurale ricorrente. Il motivo di tali scelte è però da imputare anche al diverso periodo storico degli autori, considerando che tra i due articoli è presente un gap di circa 30 anni.

2.1 Architettura

Per quanto riguarda l'architettura scelta, come già anticipato, il modello di Rumelhart e McClelland è un MLP composto da due parti:

[1] O dato un certo numero di forme flesse di un lessema, completare il paradigma: in questo caso si parla di flessione di parola

pattern associator network (rete che ha il compito di associare patterns) che apprende le diverse relazioni che si instaurano tra forma base e past-tense; questa è la parte deputata all'apprendimento

decoding network che converte la rappresentazione di *features* in rappresentazione fonologica legittima

La rete *pattern associator* è molto semplice poiché composta da soli 2 strati (uno di input e l'altro di output) senza ricorrenze e senza connessioni da un'unità di ingresso a un'altra o da un'unità di output a un'altra. Mentre il modello nel complesso (considerando sia il pattern associator che l'encoding network) prevede uno strato di ingresso che prende in input la **rappresentazione fonologica** della forma di base del verbo da apprendere, i due strati intermedi per la codifica scelta, ovvero secondo uno schema di Wickelgren (ref. 2.2.1), in particolare uno strato per la rappresentazione di *Wickelfeature* della forma base e uno per lo stesso tipo di rappresentazione del past tense; infine uno strato di uscita che restituisce in output la **rappresentazione fonologica** del *past tense* corrispondente alla forma di base. Ciascuna unità della rete è una *feature* particolare della stringa di input o output.

Analizzando invece l'architettura descritta nell'articolo di Malouf notiamo subito una maggiore numero di strati addestrabili poiché la rete, estensione dell'architettura di Thymé per flessioni nominali, è costituita da quattro *layers*: nello strato iniziale la rete prende in input un lessema, un'insieme di caratteristiche morfosintattiche e una *wordform* parziale, ovvero una lettera; questi input sono mappati (quindi concatenati tra loro) in uno strato nascosto di proiezione. Le unità di quest'ultimo strato saranno poi gli input dello strato ricorrente implementato secondo LSTM (*Long short-term memory*); arriviamo infine allo strato di output le cui unità saranno una distribuzione di probabilità di caratteri.

Nel modello di Malouf possiamo vedere che sono presenti come input iniziali una rappresentazione *localista* del contesto fonologico immediato, ovvero un modello in cui ogni elemento è identificato in modo univoco da un singolo nodo nella rete, a differenza invece della rappresentazione distribuita del modello di Rumelhart e McClelland in cui una singola parola è analizzata come un insieme di proprietà. Nel modello di Malouf la mappatura di questi input sullo strato nascosto di proiezione consente poi di trasformarli in rappresentazioni distribuite del contesto fonologico e morfologico.

2.1.1 Vantaggi della rete neurale ricorrente

Risultano evidenti le enormi differenze strutturali tra le due reti: la MLP in questo modo è sicuramente più limitante rispetto alla RNN. I maggiori vantaggi della seconda rete sono dati:

- dalla presenza di due ricorrenze necessarie per individuare dipendenze: quella più esterna (dallo strato di output a quello di input della rete) consente al modello di apprendere più facilmente *pattern* fonologici evitando così errori del tipo *wwalkkd* al posto di *walked*, mentre quella più interna (che ricorre nello strato nascosto) individua la struttura sequenziale di parole complesse
- dall'uso delle distribuzioni di probabilità dei caratteri e quindi della wordform: tale natura fa sì che il modello si possa adattare ad altre situazioni, ad esempio quando ad un lessema manca la wordform di una cella paradigmatica o, al contrario, quando ad una cella possono corrispondere più di una wordform
- dall'uso di LSTM: modello che supera i limiti del *vanishing gradient* e quindi del problema di *Long-term dependencies*, ovvero semplici RNN non sono capaci di tenere in memoria eventi

passati lontani nel tempo, quindi quando il contesto è troppo lontano; in questo modo il modello può catturare dipendenze anche distanti temporalmente

2.1.2 I vincoli dell'architettura

La rete *Multi Layer Perceptron* di Rumelhart e McClelland presenta diversi problemi di rappresentazione: a causa dei vincoli imposti dall'architettura stessa è necessario somministrare alla rete tutte parole con rappresentazione uniforme (parole della stessa lunghezza) sia in entrata che in uscita, poiché in una rete MLP la dimensione di input e output deve rimanere costante, e soprattutto ciascuna parola in input deve essere mostrata tutta in una volta sola.

Il problema di rappresentazione uniforme e statica di una parola può essere risolto attraverso il sistema ricorrente LSTM di Malouf dove le parole sono rappresentate come una sequenza temporale di eventi.

2.2 Codifica

Il lessico mentale contiene tutte le parole memorizzate dalla memoria a lungo termine e codificate come serie temporali di suoni o lettere. Secondo questo punto di vista, è necessario mantenere una codifica seriale/temporale delle parole anche quando dobbiamo affrontare il problema di come accedere ed elaborare le parole stesse.
In questo senso l'architettura connessionista ha tentato di risolvere il problema della codifica temporale utilizzando la *codifica congiunta*.

2.2.1 *Wikelcoding* di Rumelhart e McClelland

Rumelhart e McClelland a causa del limite dato dalla *pattern associator* hanno optato per il cosiddetto ***Wickelcoding***: ogni parola è rappresentata come un insieme di unità, ciascuna delle quali è legata a una particolare finestra di contesto. La lettera quindi è una tripletta di elementi: la lettera corrente insieme a quella immediatamente precedente e immediatamente successiva (CAT $\rightarrow \{\#\_C\_A, C\_A\_T, A\_T\_\#\}$). Questo tipo di codifica tuttavia rende la rappresentazione dell'input dipendente dalla lingua, di conseguenza la stessa architettura di elaborazione non può essere utilizzata per trattare più lingue che presentano vincoli differenti sul modo in cui suoni o lettere sono concatenati.
Questo limite risulta essere particolarmente problematico per quanto riguarda l'apprendimento della struttura morfologica poiché le lingue differiscono enormemente dal punto di vista strutturale (ad es. forme inglesi come *walk-walked* differiscono dall'arabo *kataba-yaktubu*) così da rendere necessario applicare strategie di rappresentazione differenti.

A causa del numero altrimenti troppo grande di unità da rappresentare secondo il Wickelcoding, Rumelhart e McClelland hanno optato per il metodo delle Wickelfeature: 16 Wickelfeature per ciascun wicklephone. In questo modo la rete ha 460 unità di output e input invece che 35^5 Wickelphones.

2.2.2 Codifica degli input in RNN

Il tipo di input e output del modello di Rumelhart e McClelland è stato fortemente criticato da Malouf in *Abstractive morphological learning with a recurrent neural network*, poiché il modello prende in input un insieme di rappresentazioni fonologiche piuttosto che forme piene.

Lo scopo principale di Malouf consiste nel generare paradigmi completi attraverso una rete neurale ricorrente. Per fare ciò si è basato sull'idea che un paradigma non sia altro che una funzione paradigmatica PF la quale mappa un lessema e un insieme di caratteristiche morfosittattiche in una parola. Dunque specificando la funzione PF si fornisce allo stesso tempo una soluzione al *Paradigm Cell Filling Problem*. L'approccio scelto è di tipo astrattivo [1] con al centro la nozione di *predicibilità*: si assume che il lessico sia costituito da un insieme di forme flesse piene a partire dalle quali è possibile astrarre sotto-costituenti.

Il sistema prende in input un lessema, ovvero una unità di base del lessico (ad es. WALK), un insieme di caratteristiche morfosintattiche e una wordform parziale, ovvero una sequenza di simboli somministrati uno alla vola (ad es. la lettera w), e restituisce una distribuzione di probabilità del carattere seguente a tale wordform. Possiamo quindi notare che, a differenza di Rumelhart e McClelland, l'input della rete non contiene informazioni sulla natura fonologica dell'output dunque la forma di base deve essere dedotta sulla base della conoscenza di altri paradigmi del dato lessico. Inoltre la rete apprende pattern di implicazione senza che ad essa vengano fornite regole sulla struttura morfologica dell'input [2].

Dal punto di vista della codifica, l'utilizzo di una RNN da parte di Malouf permette di superare alcuni fondamentali problemi di un MLP: la parola è considerata come un sequenza temporale di eventi, per cui questa verrà somministrata un simbolo alla volta ed il rientro consentirà di accumulare, a fine parola, la rappresentazione dell'intera sequenza.

I nodi nello strato di input sono:

- un vettore binario di dimensione uguale al numero di segmenti totali presenti in una lingua, più due simboli rappresentanti i confini di parola: il vettore in questione è una rappresentazione *one-hot*.
- un vettore corrispondente al lessema (1 bit) e un vettore per ciascuna cella paradigmatica (1 bit totale).

I vettori di input sono codificati come matrici dense i cui output sono poi concatenati in un *projection layer* z(t) e mappati in un layer LSTM insieme all'output calcolato nell'istante $t-1$. Le reti LSTM sono in grado di apprendere *long-term dependencies* poiché sono RNN in grado di elaborare sequenzialmente dati di input, conservando nel tempo il loro stato interno.

3 Valutazione del modello

I due modelli presentano delle differenze anche per quanto riguarda il modo di valutare la performance della rete poiché, mentre il modello di Malouf è stato valutato in termini di correttezza e capacità della rete di risolvere il compito, il modello di Rumelhart e McClelland è stato valutato sulla base del confronto tra il processo di maturazione linguistica dei bambini e il comportamento del modello.

3.1 Confronto con comportamento umano

Come già anticipato lo scopo principale di Rumelhart e McClelland consiste nel dimostrare che una rete connessionista è in grado di simulare il processo di acquisizione del past-tense da parte dei

[2] Ad esempio regole per riconoscere radice+affisso o morfemi

bambini. Si può notare infatti che inizialmente i bambini tendono ad apprendere correttamente le parole, sia che siano forme regolari o eccezioni; ad un certo punto tuttavia, le stesse eccezioni apprese durante le prime fasi di apprendimento verranno regolarizzate, non producendo più past-tense corretti (*eated* piuttosto che *ate*); infine durante l'ultimo stadio saranno in grado di produrre past-tense corretti sia per le forme irregolari che per le forme regolari. Questo tipo di andamento è chiamato *U-Shaped Learning* e, basandosi sull'apprendimento standard di un bambino, può essere riassunto nel modo seguente:

- inizialmente i bambini memorizzano semplicemente le parole (verbi per lo più ad alta frequenza e irregolari)
- deducono poi delle leggi morfologiche e ciò crea una sorta di interferenza o competizione tra le forme memorizzate e le forme generate tramite queste regole (fenomeno della regolarizzazione)
- infine i due sistemi riescono a coesistere: past-tense corretto delle forme irregolari, applicazione della forma regolare a nuove parole.

In realtà non c'è un vera e proprio cambio radicale tra la fase due e la fase tre, ma piuttosto un cambio graduale.

Il modello PDP di Rumelhart e McClelland è in grado di riprodurre una curva di apprendimento a ferro di cavallo ma ciò è possibile solo fornendo inizialmente verbi ad alta frequenza (per lo più irregolari) e poi includendo l'intero training set (quindi anche verbi regolari meno frequenti). Dunque il modello per riprodurre tale comportamento deve necessariamente ricevere porzioni differenti di verbi irregolari come input in tempi precisi.
La rete è stata addestrata utilizzando la procedura di *perceptron convergence* dato che si tratta di un problema linearmente separabile presentando come input la forma di base e come output la struttura fonologica del paste tense. A seguito di un confronto tra il past-tense calcolato e il target (past-tense corretto), si aggiustano i pesi. Il test consiste invece nel presentare alla rete la forma di base e vedere quale past-tense genera come output.

Confrontando il comportamento dei bambini e la performance del modello è possibile vedere che quest'ultimo è effettivamente in grado di riprodurre la maggior parte dei fenomeni senza fare uso di regole esplicite: è in grado di apprendere regolari e irregolari, produrre il past-tense della maggior parte dei verbi regolari più rari e riprodurre la curva di apprendimento ad U. Tuttavia sono state avanzate diverse critiche nei confronti di questo modello, esposte nel paragrafo 4.

3.2 *Performance* e *accuracy*

Malouf non intende creare un modello in grado di simulare il comportamento umano durante l'acquisizione di una lingua, ma piuttosto vuole implementare un sistema in grado di simulare una soluzione al cosiddetto *Paradigm Cell Filing problem*: data una conoscenza parziale di paradigmi per un set di lessemi, generare **correttamente** le restanti forme sconosciute.

Per poter calcolare la capacità di generalizzazione e la correttezza del modello, i paradigmi sono stati generati sulla base di lessici di forme flesse piene di 7 lingue morfologicamente complesse (finlandese, francese, irlandese, khaling, maltese, palantla, chinatec e russo). Successivamente è stata calcolata la performance della rete per ciascuna lingua attraverso *ten-fold-cross-validation*. In *k-fold-cross-validation* il dataset viene diviso in k sottoinsiemi, il metodo viene poi ripetuto k volte in cui ogni volta uno dei k sottoinsiemi viene utilizzato per valutare il test e gli altri sottoinsiemi rimanenti per il *training*.

É stata poi calcolata anche l'*accuracy* di una *baseline*, ovvero una rete più semplice rispetto al modello vero e proprio poiché deriva un paradigma completo a partire da un lemma (e non un lessema), che di per sé contiene informazioni fonetiche. L'*accuracy* ottenuta dalla *baseline* è stata poi utilizzata come stima di quanto una lingua si avvicina a un modello convenzionale di morfologia flessiva.

Le performance calcolate per ciascuna lingua sembrano essere piuttosto buone. Sulla base di tali performance prodotte, sembra che la rete sia capace di apprendere i pattern di implicazione necessari per risolvere il PCFP. I risultati della rete mostrano quindi che essa è in grado sia di imparare a generare nuove forme a partire da forme già note non segmentate, che di acquisire una conoscenza profonda per quanto riguarda la struttura interna della parola, come ad esempio individuare il tema verbale e il confine con la terminazione flessionale.

4 Conclusioni

Il modello Di Rumelhart e McClelland è stato fortemente criticato anche dallo stesso Malouf;

- l'apprendimento *U-Shaped* del modello non corrisponde a ciò che realmente accade: in realtà durante le varie fasi di acquisizione del past tense al bambino non vengono somministrati porzioni variabili di verbi regolari, ma piuttosto fra i vari stati sembra che la percentuale di verbi regolari rimanga stabile [3]
- la rappresentazione *Wickelfeature* è povera,
- il numero di errori commessi dalla rete durante la fase di test è inaccettabile: su 72 verbi regolari presentati al modello addestrato, 24 hanno generato past tense errati, quindi il 33% [3],
- non hanno considerato gli effetti semantici [3]

Inoltre la rete può solo produrre il past-tense ma non è in grado di riconoscerlo, e dunque limitata a un solo tipo di task.

Le reti neurali ricorrenti LSTM, invece, risultano essere strumenti adatti all'apprendimento di funzioni paradigmatiche: dato un lessema e un set di caratteristiche morfosintattiche, la rete è in grado di generale una *wordform* completa. Inoltre tale modello rappresenta un esempio rigoroso e formale di analisi di morfologie flessive complesse.

Riferimenti bibliografici

[1] James P. Blevins. *Word-based morphology.* Journal of linguistics, Cambridge University Press, 2006.

[2] Pirrelli V., Ferro M. e Marzi C. *Slide del corso di psicolinguistica computazionale.* 2020-2021.

[3] Pinker S. e Prince A. *On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition.* 1988.