

Una codifica lessicografica per il riconoscimento del significato delle parole utilizzando reti neurali ed algoritmi evolutivi.

Antonia Azzini, Célia da Costa Pereira, Mauro Dragoni,
Andrea G.B. Tettamanzi

Dipartimento di Tecnologie dell'Informazione,
Università degli Studi di Milano

Via Bramante, 65, 26013 Crema (CR)

Italy; Tel. +39 0373898025, Fax +39 0373898010

[antonia.azzini, celia.pereira, mauro.dragoni, andrea.tettamanzi@unimi.it](mailto:antonia.azzini,celia.pereira,mauro.dragoni,andrea.tettamanzi@unimi.it)

Abstract

L'approccio proposto riguarda l'applicazione di un modello basato su tecniche di Soft Computing per il riconoscimento del significato corretto delle parole. Per ogni parola ambigua è stato considerato il corrispondente insieme dei possibili significati ed è stato utilizzato un algoritmo evolutivo per l'ottimizzazione di classificatori definiti attraverso reti neurali artificiali.

Lo schema distribuito considerato è basato su una codifica lessicografica del contesto in cui ciascuna parola è presentata.

La validità dell'approccio è dimostrata attraverso esperimenti eseguiti su un campione rappresentativo di parole ambigue.

1. Introduzione

Il riconoscimento del significato di una parola (WSD), come indicato da una recente rassegna (Navigli 2009), consiste nell'assegnare il significato più adatto ad un particolare contesto ad una parola polisemica. Recenti studi in letteratura hanno presentato diversi approcci per l'automatizzazione del riconoscimento basati, in generale, su due passi: (i) considerare i possibili significati di una data parola, (ii) assegnare ogni occorrenza di una parola al suo significato più appropriato, in base sempre al contesto. La rappresentazione di quest'ultimo diventa quindi uno degli aspetti più importanti nel problema del WSD, e in letteratura sono stati presentati diverse modalità di rappresentazione, fra cui gli schemi distribuiti o localizzati (Cottrell 1989, Hinton 1986).

Questo lavoro propone un nuovo schema distribuito, basato sulla codifica lessicografica per rappresentare il contesto di una data

parola. Il riconoscimento del significato di una parola utilizza un approccio supervisionato basato sulla combinazione di reti neurali (NNs) e algoritmi evolutivi (EAs) (Azzini et al. 2008). L'aspetto innovativo di questo lavoro interessa l'implementazione di una nuova rappresentazione per codificare le sentenze, insieme a modifiche nell'algoritmo evolutivo.

Il progetto considera datasets di grandi dimensioni, che descrivono il contesto in cui è calato ogni significato di una parola polisemica; essi verranno utilizzati per evolvere un riconoscitore basato su NN. Viene definita quindi una classe di NNs specializzate nel riconoscimento del senso corretto della parola corrispondente.

Il contesto è definito utilizzando l'annotazione lessicografica che WordNet (C. Fellbaum 1998) assegna a ciascun significato di una parola, classificandola in una delle quarantacinque categorie basate sulla sintassi e sui raggruppamenti logici. WordNet è uno dei databases maggiormente utilizzati in letteratura in problemi legati all'analisi del testo. Esso fornisce una lista di significati per ciascuna parola contenuta, organizzati in insiemi di sinonimi (synsets).

Le attivazioni dei neuroni di ingresso delle NNs sono ottenute sommando i pattern di attivazione che codificano le parole in un dato contesto, escludendo il significato atteso (detto appunto target), rimuovendo la parola che si vuole disambiguare, le stopwords ed applicando l'algoritmo di Stemming.

Esperimenti condotti su un insieme sufficientemente rappresentativo di parole polisemiche hanno dimostrato la bontà dell'approccio. Confronti con i migliori risultati ottenuti dalle competizioni di "Semeval-1/Senseval-4" (Web ref.) hanno inoltre dimostrato come esso sia competitivo rispetto ad altre tecniche presentate in letteratura.

2. Approcci supervisionati per il WSD

Interessanti aree di ricerca nel WSD riguardano l'applicazione di approcci supervisionati usati per apprendere e classificare i significati, a partire da insiemi di dati forniti in ingresso alla fase di addestramento (training) dei classificatori stessi. In questi casi i dataset impiegati devono essere sufficientemente rappresentativi per garantire soddisfacenti risultati da parte del modello implementato anche su dati precedentemente non visti, generalmente impiegati per il test e le successive validazioni.

Uno degli aspetti più critici negli approcci supervisionati riguarda la rappresentazione del contesto in cui è usata una parola. In questo

approccio la rappresentazione è affidata agli ingressi delle reti neurali impiegate come classificatori. In questo lavoro è stata utilizzata una rappresentazione lessicografica per la definizione del contesto, basandosi sull'annotazione assegnata a ciascun significato da WordNet, classificando le informazioni in una delle quarantacinque categorie lessicografiche in base alla categoria sintattica e ai raggruppamenti logici. La Tabella 1 mostra un esempio di categorie lessicografiche.

Categoria	Descrizione
noun.artifact	nouns denoting man-made objects
noun.location	Nouns denoting spatial position
noun.process	Nouns denoting natural processes

Tabella 1: Esempio di Categorie Lessicografiche

Il contesto di ogni parola considerata w è rappresentato da un vettore di quarantacinque elementi, i contributi delle istanze di altre parole nelle frasi della categoria corrispondente. Tale contesto viene dato in ingresso alla rete neurale.

In generale il contributo $C_k(w)$ di una istanza della parola w al k -esimo componente del vettore "contesto" C è definito come segue:

$$C_k(w) = \frac{N_k(w)}{N(w)}$$

$N_k(w)$ è il numero di "synset" di w con categoria k , $N(w)$ è il numero di synset della parola w . E' possibile osservare che il contributo di una parola "monosemia" è massimo, cioè pari a 1.

Sia S la frase in cui la parola w deve essere disambiguata. Il k -esimo elemento del vettore contesto C di S è dato da:

$$C_k = \sum_{w \in S} C_k(w)$$

Partendo, per esempio, dalla frase "part aqueduct system", in cui la parola "target", *tunnel*, significa (1) "un passaggio attraverso qualcosa" e (2) "un buco creato da un animale", il contributo ad ogni input della rete neurale (C_1, \dots, C_{45}) viene calcolato come mostra la Tabella 2. Il numero di parentesi è il numero delle istanze della corrispondente categoria lessicografica nella lista dei significati per ciascuna parola.

Word	Lexicographic Category	Contribution	Word	Lexicographic Category	Contribution
part		(18 senses)	aqueduct		(1 sense)
	<noun.act>	0.110 (2)		<noun.artifact>	1.000
	<noun.artifact>	0.055 (1)			
	<noun.body>	0.055 (1)	system		(9 senses)
	<noun.cognition>	0.175 (3)		<noun.artifact>	0.111 (1)
	<noun.communication>	0.055 (1)		<noun.attribute>	0.111 (1)
	<noun.location>	0.055 (1)		<noun.body>	0.222 (2)
	<noun.object>	0.055 (1)		<noun.cognition>	0.334 (3)
	<noun.possession>	0.055 (1)		<noun.group>	0.111 (1)
	<noun.relation>	0.055 (1)		<noun.substance>	0.111 (1)
	<verb.motion>	0.110 (2)			
	<verb.social>	0.055 (1)			
	<verb.contact>	0.110 (2)			
<adv.all>	0.055 (1)				

Tabella 2: Input della frase “part aqueduct system”

3. L'algoritmo neuro-evolutivo

La progettazione delle reti neurali usate come classificatori è basata su un approccio evolutivo, validato su differenti benchmark e problemi reali (Azzini Tettamanzi (2008), Azzini et al. (2008)). La popolazione degli individui, i classificatori, è definita attraverso perceptron multi strato (MLPs), un tipo di NN feedforward, soggette ad addestramento, evolute attraverso l'ottimizzazione congiunta delle topologie e dei pesi.

Ogni rete è definita con un numero prestabilito di neuroni in ingresso, corrispondente alla dimensione del vettore di contesto; la dimensione dell'uscita di una rete è data dal numero di significati della parola target. Variano invece il numero di livelli intermedi di una NN e il numero di neuroni definiti in ciascun livello, per poter garantire una diversità di individui nella popolazione. Il numero di neuroni in un livello deve però essere sempre maggiore od uguale al numero di quelli del livello precedente, per evitare problemi con strutture “a clessidra”, causa di diminuzione delle prestazioni.

Ad ogni generazione la metà migliore degli individui viene selezionata con un algoritmo di troncamento sul totale degli individui, viene duplicata per definirne nuovi a sostituzione dei peggiori, ed infine viene permutata casualmente. L'elitismo impedisce che l'individuo migliore in una popolazione venga cancellato di volta in volta. Successivamente, vengono mutati i pesi e la topologia degli

individui della popolazione, vengono riaddestrate le reti neurali e viene calcolata infine la funzione di fitness, funzione obiettivo dell'algoritmo evolutivo, per ciascun individuo, salvandone il migliore con le informazioni statistiche dell'intero processo evolutivo. La fitness di un individuo si basa sulla matrice di confusione, in particolare sulla differenza fra il numero di neuroni di output e il parametro che corrisponde alla somma degli elementi della diagonale della matrice di confusione normalizzata per righe. La matrice di confusione è uno strumento impiegato per visualizzare il rapporto fra i casi riconosciuti in modo corretto e quelli non riconosciuti, rispetto all'insieme totale dei dati considerati, con i rispettivi valori attesi (target).

L'intero algoritmo evolutivo non usa l'operatore di ricombinazione, a causa degli effetti negativi che ne derivano (Yao, Liu (1997)).

4. Esperimenti e risultati

La campagna di esperimenti è stata realizzata utilizzando gli stessi settaggi utilizzati nel precedente lavoro (Azzini, Tettamanzi (2008)). I risultati degli esperimenti sono riportati nella Tabella 3. In 9 casi su 15 la codifica lessicografica migliora i risultati ottenuti con la codifica posizionale, mentre in altri 3 casi i risultati sono molto simili.

Per analizzare più in dettaglio le prestazioni del sistema vengono presentate in Figura 1 le matrici di confusione relative ai risultati ottenuti dal sistema su ciascuna delle parole analizzate.

Nei casi di *bar*, *channel*, *circuit*, *detention*, *grip*, *hearth*, *material*, *memory* e *post* l'algoritmo riesce a distinguere in modo soddisfacente la maggior parte dei significati.

Non si può dire lo stesso per le parole *bum*, *day*, *holiday* e *lady* dove i risultati ottenuti sono conseguenza della scelta dell'algoritmo neuro evolutivo di riconoscere sempre il significato più probabile.

Un caso particolare invece è dato dalla parola *dyke*, nella quale l'algoritmo riesce a distinguere con una buona precisione entrambi i significati. Questo risultato dimostra come, in presenza dataset sufficientemente rappresentativi del problema e con significati ben distinti, l'approccio presentato sia in grado di produrre soddisfacenti risultati.

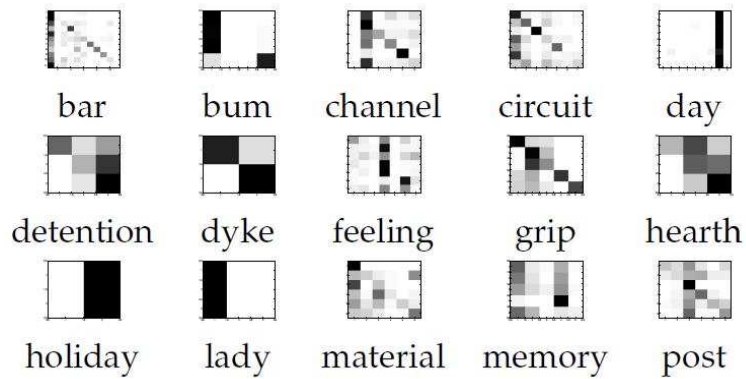


Figura 1: Matrici di confusione delle parole analizzate.

Word	Expected Accuracy	Letter Count Accuracy		Positional Accuracy		Lexicographic Accuracy	
		ANNs	EANNs	ANNs	EANNs	ANNs	EANNs
bar	-	34.97	33.93	34.47	29.57	35.15	44.02
bum	57	42.39	42.64	42.45	47.12	62.62	63.94
channel	43	35.88	32.89	35.88	39.75	62.91	65.90
circuit	35	47.04	42.16	47.01	50.43	47.01	62.67
day	35	28.69	35.26	28.98	29.46	29.04	28.24
detention	75	88.25	92.34	68.68	95.19	70.18	71.86
dyke	75	96.20	97.40	91.67	98.73	95.83	96.38
feeling	35	30.26	32.06	30.26	35.38	28.16	40.84
grip	52	42.51	48.93	41.98	56.77	57.93	63.19
hearth	62	47.67	50.88	47.67	61.94	57.09	59.03
holiday	75	93.87	88.56	88.92	90.75	93.92	93.23
lady	62	45.42	46.61	45.82	50.15	45.82	45.82
material	43	68.35	68.36	68.36	69.73	68.37	75.62
memory	52	46.60	57.12	46.60	62.00	53.77	63.55
post	43	59.31	59.05	53.38	59.70	60.16	71.97

Tabella 3: Risultati ottenuti dall'algoritmo di disambiguazione e confronto con gli esperimenti condotti in precedenza (Azzini Tettamanzi 2008).

5. Conclusioni

A prima vista, la creazione di una singola rete neurale per ciascuna parola ambigua potrebbe sembrare un obiettivo irrealizzabile. Tuttavia, sono presenti in WordNet solamente 15,935 parole ambigue.

Dagli esperimenti condotti risulta che la realizzazione di una rete neurale richiede circa due ore di tempo computazionale, ipotizzando quindi di utilizzare circa 30 personal computer a ciclo continuo, in circa 45 giorni tutte le parole avranno associata la loro rete neurale.

La stima di occupazione di memoria di tutte le 16.000 reti neurali è di circa 30 Mbytes e l'utilizzo della rete neurale per il corretto riconoscimento del significato di una frase richiede pochi millesimi a livello computazionale. Ciò significa che un tale approccio potrebbe essere applicato nella realtà al problema del riconoscimento dei significati corretti con risultati accettabili sia in termini di tempo che di risorse computazionali necessarie.

Bibliografia

1. Azzini, A., Dragoni, M., da Costa Pereira, C., Tettamanzi, A.G.B.: Evolving neural networks for word sense disambiguation. Proc. Of HIS'08, pp. 332-337, 2008.
2. Azzini, A., Tettamanzi, A.G.B.: Evolving neural networks for static single-position automated trading. J. of Artificial Evolution and Applications, pp.1-17, 2008.
3. Yao, X., Liu, Y.: A new evolutionary system for evolving artificial neural networks. IEEE Trans. On Neural Networks, vol.8, pp. 694-713, 1997.
4. R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1-69, 2009.
5. G. Cottrell. *A Connectionist Approach to Word Sense Disambiguation*. Pitman, London, 1989.

6. G. Hinton, J. McClelland, and D. Rumelhart. Distributed representations. In *Parallel Distributed Processing: explorations in the microstructure of cognition*. MIT Press, Cambridge, MA, 1986.
7. C. Fellbaum. WordNet: An Electronic Lexical Database, MIT Press, 1998.
8. Web reference: "www.senseval.org"