10.12871/CLICIT2014154

Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare.

Maria Palmerini
Cedat 85
m.palmerini@cedat85.com

Renata Savy
DipSUm / Lab.L.A. Università di Salerno
rsavy@unisa.it

Abstract

Italiano. *Il lavoro presenta i risultati di un lavoro di classificazione e analisi linguistica degli errori di un sistema di riconoscimento automatico (ASR), prodotto da Cedat'85. Si tratta della prima fase di una ricerca volta alla messa a punto di strategie di riduzione dell'errore.*

English. *The research project aims to analyze and evaluate the errors generated by Cedat 85's automatic speech recognition system (ASR), in order to develop new strategies for error reduction. The first phase of the project, which is explored in this paper, consists of a linguistic annotation, classification and analysis of errors.*

1 Introduzione

Il progetto di ricerca è nato da una collaborazione fra l'Università di Salerno e Cedat 85, azienda leader in Italia nel settore del trattamento automatico del parlato. Lo scopo del progetto è una valutazione accurata degli errori prodotti da un sistema di trascrizione automatica del parlato (ASR), passati al setaccio di una più fine analisi linguistica e successiva metadatazione.
La stima più utilizzata del *word error rate* (WER) di un sistema ASR è calcolata in maniera automatica e si basa sull'analisi di una trascrizione manuale (allineata al segnale) e la relativa trascrizione ottenuta dal sistema ASR. Su questo confronto vengono individuate le parole errate (*Substitutions*), quelle mancanti (*Deletetions*) e quelle erroneamente inserite (*Insertions*) nonché le parole totali (N) per una valutazione:

$$WER = \frac{(S+D+I)x100}{N}$$

Questa stima non entra nel merito della causa né della rilevanza dell'errore, costituendo piuttosto un riferimento di massima per una valutazione grossolana di un sistema ASR, senza alcuna indicazione sulla sua reale utilità e adeguatezza, né sulle possibilità di intervento e miglioramento.
Gran parte dei sistemi ASR di ultima generazione, che lavorano su parlato spontaneo, utilizzano tecnologie ed algoritmi che possono sfruttare al meglio l'enorme potenza di calcolo attualmente disponibile, ma differiscono in modo rilevante nella scelta dei parametri, dei passi intermedi, nei criteri di selezione dei candidati più probabili, negli strumenti per il trattamento dei dati di addestramento. Un criterio 'qualitativo', oltre che quantitativo, di valutazione degli errori si rende necessario per un adeguamento del sistema all'ambiente di riferimento, e per l'indicazione su eventuali interventi migliorativi.
Studi recenti, sia di ambito tecnologico che linguistico e psicolinguistico, indicano correlazioni tra errori e frequenza nel vocabolario o nell'uso delle parole, velocità d'eloquio, ambiguità (omofonia) e confondibilità acustica (coppie minime e sub-minime). Mancano tuttavia studi sistematici che prendano in considerazione la correlazione con classi morfo-lessicali, strutture fonologiche e sillabiche, sequenze sintagmatiche, ordine dei costituenti e soprattutto, fattori prosodici.
In questo contributo presentiamo una prima parte dei risultati di una ricerca più ampia sul peso di questi fattori, soffermandoci sui criteri della classificazione linguistica dei dati e sulle correlazioni ottenute tra presenza (e tipo) di errore e categorie fono-morfologiche e morfo-sintattiche.

2 Corpus e metodo di analisi

Cedat 85 ha messo a disposizione un corpus di registrazioni audio (che chiameremo *test set*, v. §2.2) con relative trascrizioni manuali e trascrizioni prodotte automaticamente dal proprio sistema ASR. Su questi dati è stato calcolato il *word error rate* (WER) in modo automatico, grazie al tool Sclite, componente del pacchetto *Speech Recognition Scoring Toolkit* (SCTK) realizzato dal *National Institute of Standards and Technology* (NIST).
Sono inoltre stati messi a disposizione il *phone set* e il dizionario utilizzati dal sistema ASR.

2.1 Il sistema ASR

Il sistema per il riconoscimento automatico del parlato continuo di Cedat 85 è un sistema di ultima generazione, *speaker independent* (che quindi non richiede addestramento specifico sulla singola voce), basato su modelli statistici di tipo markoviano[1]. Nel sistema ASR analizzato la decodifica del parlato avviene grazie a due moduli che interagiscono fra loro: un 'modello acustico', deputato al riconoscimento dei suoni significativi all'interno del segnale, e un 'modello di linguaggio', cui spetta l'individuazione di parole singole (unigrammi) e sequenze di parole (bigrammi e trigrammi). Entrambi i moduli si basano su un dizionario (lessicale e fonologico). I modelli acustici per la lingua italiana sono stati addestrati su centinaia di ore di parlato proveniente da vari ambienti sia microfonici, sia telefonici. Sono stati messi a punto diversi modelli di linguaggio, dal politico al televisivo, dalle lezioni universitarie al giudiziario.

2.2 Il test set

Il *test set* sottoposto ad analisi è suddiviso in 4 *subset* appartenenti a 4 diversi domini; 3 di tipo microfonico (politico, televisivo, giudiziario) e uno di tipo telefonico (sms vocali e telefonate di call center). I subset microfonici ammontano a circa 25min. di parlato ognuno, mentre il subset telefonico è composto da 109 messaggi vocali e 20 min. circa di interazioni di call center.
Su tale *test set* è stato calcolato il WER, suddiviso nelle tre categorie di errori: *Insertion* (I), *Deletion* (D), *Substitution* (S).

2.3 Metodo di classificazione

L'indagine è stata svolta in 3 fasi. Nella fase preliminare le categorie del WER sono state scorporate sui 4 diversi domini.
Nella seconda fase si è proceduto alla catalogazione degli errori per ogni dominio secondo il sistema di metadatazione linguistica (descritto in §3). L'analisi uditiva è stata corredata da una minima osservazione spettrografica. Per ciascuna stringa è stato effettuato il confronto puntuale tra le due trascrizioni per ogni item marcato da errore; l'etichettatura ha riguardato sempre l'elemento del *Reference text* (trascrizione manuale), fatta eccezione per i casi di 'inserzione' in cui è stato marcato l'elemento inserito dal sistema automatico. A valle dell'etichettatura, sono stati scorporati dal WER tutti i casi di 'falso errore', attribuibili a incomprensione o refusi del trascrittore umano. Il calcolo delle correlazioni riguarda quindi il corpus 'epurato'.
Infine, in una terza fase è stato effettuato un *PoS-tagging* di tutti i testi di riferimento dei 4 subset, allo scopo di 'pesare' i dati delle correlazioni individuate tra errore e categorie lessicali e ricavare indicazioni più puntuali e impiegabili per future ottimizzazioni del modello.

3 Il sistema di annotazione

Il modello di annotazione linguistica è stato progettato dal Laboratorio di Linguistica Applicata dell'Università di Salerno, mettendo a punto un sistema di metadatazione che prende in esame diverse caratteristiche. Schematicamente possiamo distinguere tra tre tipi di categorizzazione: 1) lessicale (*Pos*), ulteriormente articolata al suo interno; 2) 'morfologica' (implicata esclusivamente per alcune PoS); 3) 'fonetico-fonologica'. Di seguito si presenta l'elenco delle categorie del modello e relativi valori che ognuna può assumere. Tutte le etichette si riferiscono alle parole grafiche (unigrammi) considerate dal sistema.

Error Type: indica il tipo di errore secondo il sistema di misurazione automatica; può assumere i valori di I(nsertion), D(eletion), S(ubstitution).

Error Category: indica la categoria lessicale della parola oggetto dell'errore; assume i valori di Noun (N), Verb (V), Adjective (Adj), Adverb (Adv), FunctionWord (FW) and Other (O); quest'ultima categoria marca fenomeni di disfluenza, ripetizioni, false partenze e simili.

Error Subcategory: prevede una sottocategorizzazione sintattico-semantica delle PoS maggiori e una capillare descrizione delle parole funzionali, delle esitazioni e altri fenomeni (*marcatori discorsivi, false partenze, autocorrezioni, ripetizioni, pause piene, lapsus, errate pronunce*).

Verb + Clitics: assume valore 'True' (T) nel caso in cui il target dell'errore sia una forma verbale con clitico pronominale (es: *dimmi*).

Derivate: indica se il target dell'errore in questione è una parola derivata, e quindi presenta affissazione; i valori possibili per questo campo sono 'P','S' e 'P+S'.

Position: riferisce la posizione di Avverbio rispetto a Verbo e Aggettivo rispetto a Nome; assume valori 'Pre' e 'Post'.

Morphological Complexity: indica il grado di composizione morfologica della parola target secondo una 'scala di morfo-complessità' calcolata partendo dal lessema-base e aggiungendo +1 per ogni nuovo morfema, ad esempio:

industria	1
industri-ale	2
industri-al-izzare	3
industri-al-izza-zione	4
de-industri-al-izza-zione	5

[1] Il sistema è attualmente impiegato in numerose applicazioni e servizi già commercializzati da Cedat 85.

Phonological Length: indica la lunghezza in fonemi del target di errore, basata sulla trascrizione fonologica del vocabolario di riferimento.

Syllabic Length: indica la lunghezza in sillabe fonologiche del target di errore.

Accentual Type: indica il tipo accentuale del target di errore: tronco, piano, sdrucciolo, bisdrucciolo.

Omophones: indica la possibile esistenza di omofoni per la parola target; assume valori booleani (t/f).

Minimal Pairs: indica la possibile esistenza di coppie minime con la parola target; assume valori booleiani (t/f).

Alcune delle categorie sopra elencate presentano evidenti correlazioni in partenza: la presenza di clitico pronominale sul verbo o di affissazione, ad esempio, implica complessità morfologica e può comportare maggiore lunghezza fonologica e sillabica, nonché influenzare il tipo accentuale. Ciononostante, ogni parametro è stato valutato separatamente, per poter *a posteriori* verificare la concomitanza di più fattori critici.

4 Primi risultati

In questa prima analisi dei risultati riportiamo solo le correlazioni rivelatesi significative e soprattutto adeguate ad avanzare ipotesi utili per indirizzare le indagini successive. I valori nelle tabelle si intendono come percentuali sul totale degli errori del corpus di controllo.

La prima verifica linguistica riguarda la distribuzione dell'errore nelle diverse categorie lessicali, che mostra una situazione omogenea, diversa solo per il dominio telefonico.

	N	V	ADJ	ADV	FW	O
politico	11,2	11,6	5,1	3,8	**29,3**	**38,7**
media	15,8	18,7	2,5	3,2	**25,7**	**34,2**
giustizia	7,7	17,7	2,6	3,5	**33,2**	**35,2**
telefonico	17,6	21,4	3,6	8,1	**33,8**	15,3

Tabella 1. Distribuzione di Error category nei 4 subset.

I dati in tab.1[2] evidenziano una pesante concentrazione dell'errore per la classe delle parole funzionali (FW) e delle produzioni disfluenti (O), oscillante tra il 30 e 38%. Tra le parti variabili del discorso sono scarsamente affetti da errore aggettivi e avverbi (fatta eccezione per il corpus telefonico), mentre una percentuale leggermente più alta si registra nella classe dei e, per i corpora TV e Telefonico, anche per la classe dei nomi.

I successivi dati significativi ci sembra riguardino la correlazione tra percentuale di errore e complessità morfologica, sillabica e fonologica (le ultime valutate in termini di 'lunghezza'). Le tabb.2 e 3 riportano in dettaglio i dati delle prime due categorie (mentre è più difficile riassumere i dati sulla lunghezza fonologica, altamente variabile e disomogenea):

	0	1	2	3	4	5
politico	**39,1**	**38,7**	21,4	0,8	-	-
media	**29,6**	**51,8**	14,1	3,9	0,7	-
giustizia	**35,2**	**34,3**	**24,4**	5,4	0,6	-
telefonico	10,3	**42,2**	**38,0**	9,2	0,2	0,2

Tabella 2. Distribuzione del WER nella categoria Morpho_complex dei 4 subset (con valore di morfocomplessità 0 sono state indicate le esitazioni e i fenomeni di disfluenza)**.**

Appare netta, dunque, un'elevata concentrazione di errori per le parole a bassa complessità morfologica (0-2), mentre quasi nulla per parole con valore di complessità morfologica superiore a 5.

	1	2	3	4	5	6
politico	10,2	**33,9**	**28,8**	13,6	10,2	3,4
media	11,1	**35,6**	17,8	15,6	**20,0**	-
giustizia	8,3	**33,3**	**38,9**	8,3	8,3	-
telefonico	14,3	**43,9**	**26,5**	9,2	4,08	2

Tabella 3. Distribuzione del WER nella categoria Syllabic length dei 4 subset.

In ultimo, sembra emergere una tendenza dell'errore (con poche eccezioni) a diminuire in modo direttamente proporzionale all'aumentare della lunghezza della parola: le parole bi- e trisillabiche concentrano, in media, oltre il 30% di errori per tutti i corpora; solo le parole monosillabiche contrastano questa tendenza generale. I dati sulla lunghezza fonologica indicano più affette da errore le parole costituite da 1 a 5 fonemi (fin oltre il 60% per quelle monofonematiche).

L'errore, dunque, si concentra sulle parole di lunghezza medio-bassa e a ridotta complessità morfologica, per ridursi poi in modo significativo nelle parole più complesse e più lunghe. Le due categorie PoS maggiormente affette da errore di riconoscimento (FW e O) sono, infatti, anche quelle che correlano con bassi o nulli valori di complessità morfologica e numero di fonemi.

Un ulteriore conteggio si rende però necessario per valutare il peso e l'incidenza del WER sulle diverse categorie lessicali. In tabella 4 riportiamo i dati di frequenza delle diverse PoS rispetto all'intero corpus, mentre in tabella 5 le percentuali di errore ricalcolate su questo insieme:

	N	V	ADJ	ADV	FW	O
politico	**23,3**	14,9	10,3	7,5	**35,8**	8,2
media	**28,5**	15,9	9,2	6,5	**36,7**	3,1
giustizia	**20,3**	**20,3**	7,0	9,8	**36,3**	6,2
telefonico	**21,7**	19,5	6,9	13,1	**31,2**	7,7

Tabella 4. Dati del PoS tagging su tutte le parole dei 4 subset.

	N	V	ADJ	ADV	FW	O
politico	7,3	11,8	7,6	7,7	12,4	**73**
media	4.0	8,5	2,0	3,5	5,1	**82,3**
giustizia	5,5	12,4	5,2	5,0	13,0	**83,3**
telefonico	**27,3**	**38,8**	**17,5**	**20,7**	**36,4**	**66,2**

Tabella 5. Incidenza dell'errore ricalcolata sul totale delle parole del corpus divise in categorie.

[2] Le tendenze regolari sono segnalate in grassetto, mentre le celle ombreggiate evidenziano dati in controtendenza.

Le PoS maggiormente affette da errore (FW e O, tab.1) hanno distribuzione frequenziale molto diversa nel corpus (tab.4): le prime, com'era prevedibile, mostrano un alto numero di occorrenze (con frequenza >30%, direttamente seguite dai Nomi); le seconde, invece, sono poco frequenti rispetto al totale delle parole del *test set* (solo il 3-8%). Ne deriva che l'incidenza dell'errore (tabella 5) è molto più significativa nel secondo caso, raggiungendo livelli anche molto maggiori dei 2/3 degli items (tra il 66 e l'83% del totale).

5 Considerazioni preliminari

Sebbene i risultati sopra esposti rappresentino un'elaborazione parziale dei dati dell'analisi del WER condotta nella ricerca, essi consentono di avanzare alcune considerazioni preliminari a future e più approfondite valutazioni.

In primo luogo, volendo misurare globalmente l'efficienza del sistema di trascrizione basato su ASR, occorre interpretare i dati inclusi in tabella 5, che mostrano percentuali di errore basse o trascurabili, comprese tra il 2% e il 13%, equamente suddivise per tutte le PoS. Fa eccezione il dominio 'telefonico' (per il quale v.oltre). Se una buona parte del WER complessivo (>25%) incide sulla categoria delle FW di un testo (tab.1), è pur vero che essa ha valori di frequenza altissimi che normalizzano l'incidenza dei mancati riconoscimenti del sistema, rendendola comparabile ad altre PoS, nonostante la loro minore complessità morfologica ed estensione fonologica.

Questo dato è d'altronde coerente col funzionamento del sistema ASR, nel quale agiscono, compensandosi, il modello acustico, che riconosce con maggiore accuratezza parole dotate di maggior 'corpo fonico', e il modello di linguaggio, che fornisce miglior supporto sulle stringhe di parole più ricorrenti, riuscendo ad integrare il riconoscimento di parole grammaticali dove l'informazione acustica è più carente (anche per fenomeni di coarticolazione e ipoarticolazione).

Una valutazione diversa va riservata ai Nomi, che mostrano un comportamento parzialmente oscillante: concentrano, infatti, percentuali variabili del WER (tab.1), anche se la loro incidenza appare normalizzata nel rapporto tra loro frequenza assoluta (22-28% sull'intero corpus) e i casi di mancato riconoscimento (tra il 4 e il 7%). In ogni caso, come classe aperta, essi sono in genere meno prevedibili e maggiormente specifici rispetto a ciascun dominio: richiedono pertanto una massiccia 'personalizzazione' del vocabolario (implementazione effettuata con addestramento sullo specifico dominio), più semplice su alcuni domini a lessico meno variabile (politico e giudiziario), più aleatoria su domini più liberi.

Risulta così che un'incidenza davvero significativa del WER si ottiene unicamente nella classe etichettata come O(ther) che racchiude in genere fenomeni di disfluenza del parlato costituiti da espressioni non lessicali, esitazioni, parole interrotte o mal pronunciate; elementi non inclusi nel vocabolario né considerati nel modello di linguaggio e quindi soggetti a errori di riconoscimento quasi per *default.* Va considerata, inoltre, l'alta variabilità delle possibili forme che essi assumono nella trascrizione ortografica manuale, dove è inevitabile un elevato tasso di interpretazione e resa grafica soggettiva, in mancanza di un modello di trascrizione standardizzato. Dal confronto tra queste rese variabili e il tentativo del sistema ASR di associarle ad entrate del vocabolario acusticamente più 'vicine' deriva l'alto tasso di WER ad esse associato (>35% del WER complessivo, >66% sul totale delle occorrenze).

A parte quest'ultimo dato, dunque, l'errore non sembra essere correlato significativamente a particolari categorie lessicali, quanto piuttosto all'estensione e al 'corpo' delle parole: unità lessicali più estese, infatti, contengono maggiori informazioni acustiche e devono competere con un minor numero di candidati simili.

6 Conclusioni e sviluppi successivi

A valle di questa preliminare fase di analisi ci sembra si possa azzardare una prima conclusione importante: la valutazione quantitativa del *word error rate* sovrastima le falle di riconoscimento di un sistema ASR. La metadatazione linguistica effettuata e la successiva valutazione qualitativa normalizza i dati del WER e reindirizza la maggior quota verso fenomeni non lessicali, imprevedibili quanto poco significativi per la misura dell'efficienza del sistema. In quest'ambito, oltretutto, l'indecisione e la confusione di resa grafica sono pressoché pari per la trascrizione automatica e quella manuale. Ciò nonostante, il peso degli errati riconoscimenti di questi segmenti può essere ridotto adottando uno schema di annotazione più fine, sia in termini di norme più salde per i trascrittori, sia come modello per il sistema ASR. Ci limitiamo infine a ipotizzare che alcuni secondari interventi sul *phone set*, l'arricchimento del vocabolario con le varianti fonetiche possibili, e un migliore trattamento dei fenomeni prosodici potrebbero migliorare di qualche grado le prestazioni del sistema.

References

Daniel Jurafsky, James H. Martin. 2009. *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, second edition, New Jersey, Pearson, Prentice Hall.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003). Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. *IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, US Virgin Islands.

Patti Price. 1990. Evaluation of Spoken Language System: the ATIS domain. *Proceedings of DARPA Speech and Natural Language Workshop*, Hidden Valley, PA.