Improving machine learning approaches to (sub)classification of Primary Progressive Aphasia using fine-grained linguistic features

Cristiano Chesi[1], *Eleonora Catricalà*[2], *Antonio Miozzo*[3], *Alessandra Marcone*[3], *Stefano Cappa*[1]

[1]Center for Neurocognition and Theoretical Syntax (Ne.T.S.), IUSS Pavia

[2]Istituto Neurologico Carlo Besta, [3]Università degli Studi di Brescia,

[3]Istituto di ricovero e cura a carattere scientifico Ospedale San Raffaele

The proposal, in short. Machine Learning approaches can perform a successful classification of Primary Progressive Aphasia (PPA) variants (Garrard et al. 2013). The accuracy of these methods for PPA sub-classifications is promising, also in very sparse contexts of connected speech productions (picture description elicitation task, generating speech samples smaller than 100 tokens). This result has been obtained by including highly informative phonetic, morpho-syntactic and semantic feature information, mainly consisting of phoneme frequency, (bi)-syllabic repetition patterns, out-of-vocabulary term frequency, cues for syntactic truncated structures and characterizing low-frequency content word distribution.

Background. Purely statistic "bag-of-words" approaches to text classification have been proved to be sufficiently accurate in distinguishing transcribed speech samples along many clinical dimensions (Garrard et al. 2010). Computationally simple stochastic models (e.g. Naive Bayes Multinomial, NBM) have been used to build decision trees, simply relying on raw word frequency information. These methods attaint at a very good level of discriminatory power, for instance, between patients diagnosed with Semantic Dementia (SD) and normal controls (NC) matched by age (Garrard et al. 2013). The classifiers trained on feature vectors built on token frequency reached a classification performance above 90%. Also finer grained sub-classification of SD patients (distinguishing between right- vs. left-temporal predominant atrophic patterns) achieved a significant level of accuracy close to 90%. The latter result is obtained by reducing the feature vector to highly significant distinctive features: i. low frequency content words, ii. generic terms, iii. components of metanarrative statements.

Standard classifications in PPA. PPA is a language specific disorder associated with atrophy of frontal and temporal regions, primarily in the left hemisphere (Mesulam 1982). PPA principal sub-types are three (Gorno-Tempini et al. 2011): non-fluent/agrammatic (gPPA), semantic (sPPA) and logopenic/phonological (lPPA). gPPA is characterized by clear agrammatism in language production, effortful, halting speech with phonological errors and distortions and, usually, impaired comprehension of complex sentences even though single-word comprehension and object knowledge is preserved; sPPA is diagnosed mainly in presence of impaired object naming and single-word comprehension, often associated with difficulties with low-frequency, low-familiar objects knowledge and spared repetition and speech production; lPPA, presents both impaired single-word retrieval and repetition of sentences and phrases, frequently coupled with phonological errors in speech and naming, no major loss on single word comprehension and object knowledge and no frank agrammatism.

Comparative assessment using simple vector models. The corpus used to test our classifiers consisted of 13 elicited samples from patients with an initial PPA diagnosis and 6 samples of speech from normal controls (NC). The picnic picture description test (Western Aphasia Battery, Kertesz 1982) was used for elicitation. The productions have been transcribed using standard orthography whenever possible. The corpus consistency is 2256 words/tokens and 488 word forms/types.

Below a summary of some of the main speech characteristic dimensions used for diagnosis:

subjects	*A*	*B*	*C**	*D*	*E*	*F*	*G*	*H**	*I*	*J*	*K*	*L*	*M*
WPS	0,9	1,4	0,8	0,2	0,5	0,5	1	1,1	0,9	1,2	1,5	0,7	1
nouns	29	31	24	8	14	21	12	15	16	21	19	21	11
MF	81,3	208,8	126,6	30,50	104,8	70,7	111,9	78,5	28,3	432,3	41,1	134,8	171,6
diagnosis	gPPA	sPPA	gPPA	gPPA	lPPA	gPPA	gPPA	gPPA	lPPA	sPPA	gPPA	sPPA	sPPA

Table 1. production speed (words per second, *WPS*) number of nouns produced (*n*) mean nouns frequency (*MF*). *Initially classified as gPPA, then excluded.

At first, following Garrard et al. (2013), we used a feature vector composed by all available token frequencies (with neither lemmatization nor frequency normalization) to test Information Gain (IG, Mitchel 1997): IG synthetic value is a measure of the discriminability efficiency of each feature: higher values indicate that the feature correctly characterizes only one subset, while lower IG values indicate lower capacity to discriminate among subclasses since the feature is equally distributed among them. As expected, only few features presented high IG (e.g. 'molo': 0.609, 'sfondo': 0.498, 'sponda': 0.498, 'laghetto': 0.498, 'donna': 0.466, 'parcheggiata': 0.403 …). We used only 40 features among these (≅ 8% of the original set) to train our classifiers using NMB models, and we obtained a significant discrimination between PPA and NC (.89) ($p < 0.5$). The same feature vector scored less well on sub-discrimination (gPPA vs. sPPA vs. lPPA) (.77, $p < 0.5$ only for some comparisons).

Improving the discriminative power using richer linguistic information. In addition to the 20 lexical frequency features we tested other dimensions to improve PPA sub-varieties discriminability:

a. Raw character frequency (character distribution roughly correlates with phonemes usage; we might expect apraxia of speech to be representable by these features);
b. Bi/Tri-grams duplication, without repetitions (disfluencies are characterized by repetition of segments that often match the syllabic level; this feature also counts pauses, expressed by sequences of dots, and long hesitations "emmm emm");
c. Out-Of-Vocabolary tokens (we used Morph-it lexicon, Zanchetta et al. 2005, to classify OOV words);
d. Truncated structures (sequences terminating by functional, closed-class, words, like determiners or complementizers: in the first case we aim at counting hesitations before nouns retrieval; in the second, we expect to isolate complex context like headed relative clauses, e.g. "il bambino *che* …")

All these features show high IG values and a better fitting performance with respect to the three subclasses under analysis (.92 accuracy, $p < 0.5$ in all conditions).

Discussion. Far from being a fully automatic classification method, we showed that Machine Learning (naïve) approaches can be improved using richer linguistic features and their accuracy might help clinicians in tracking, as precisely as possible, worsening or improving in pathological speech productions.

References

Garrard & Forsyth (2010) *Neurocase*, 16: 520-528
Garrard, Rentoumi, Gesierich, Miller & Gorno-Tempini (2013) *Cortex*
Kertesz (1982) Western Aphasia Battery. New York: Grune and Stratton
Mesulam (1982) *Annual of Neurology*, 11:592–598
Mesulam, Wieneke, Thompson, Rogalski & Weintraub (2012) *Brain*, 135(5), 1537-1553
Mitchell (1997) *Machine Learning*. McGraw-Hill
Zanchetta & Baroni (2005) *Corpus Linguistics*, 1(1).