Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities

Laura Franzoi[1,3], Andrea Sgarro[2,3]
[1]University of Bucharest (Ro), [2]University of Trieste (I)
[3]Human Language Technologies Research Center (Ro)

KES 2017
Marseille, september 2017

Contents

A reminder:

distances or dissimilarities $d(x, z)$
between objects (strings) $x, y \in \mathcal{A}$

Shannon's distinguishability $\delta(x, y)$ is the combinatorial center

$$\delta(x, y) \doteq \inf_{z \in \mathcal{A}} \left[d(x, z) \vee d(y, z) \right]$$

often, as happens here, the *infimum* is an actual *minimum*

send either input x or input y through a *noisy channel* and decode the output z by minimum dissimilarity (minimum distance)

Criterion:

$\delta(x, y) = w \iff$ the decoder corrects all errors of weight $d(\ , z) < w$

send either input x or input y through a *noisy channel* and decode the output z by minimum dissimilarity (minimum distance)

Criterion:

$\delta(x, y) = w \iff$ the decoder corrects all errors of weight $d(\ , z) < w$

Caveat:

in such a general criterion one needs
both dissimilarities (distances) and distinguishabilities

Abstract fuzzy metric spaces, at least as we need them

i) $0 \leq d(x,x) \leq d(x,y)$, $d(x,y) = 0 \implies x = y$
ii) $d(x,y) = d(y,x)$
iii) $d(x,z) + d(z,y) \geq d(x,y)$

Objects x with positive "self-distance" $d(x,x)$, or positive *fuzziness* $f(x)$ are called fuzzy, else they are crisp

Defuzzification is quite easy,
just impose $d(x,x) = 0$, but doesn't always pay

For fuzzy and crisp metric distances:

Basic bounds on distinguishability:

$$\frac{d(x,y)}{2} \leq \delta(x,y) \leq d(x,y)$$

the optimizing $\underline{z}$ in $\delta(\underline{x}, \underline{y})$ is as near as possible to both $\underline{x}$ and $\underline{y}$

Euclidean insight: $x\_\_\_\_\_z\_\_\_\_\_y$ $\underline{z}$ is half way

for integer metric distances of practical interest in coding, e.g. Hamming distance or edit distance, the lower bound is "almost" achieved

$$\delta(\underline{x}, \underline{y}) = \left\lceil \frac{d(\underline{x}, \underline{y})}{2} \right\rceil$$

which seems to make the notion of distinguishability useless

the optimizing $\underline{z}$ in $\delta(\underline{x}, \underline{y})$ is as near as possible to both $\underline{x}$ and $\underline{y}$

Euclidean insight: x_____z_____y $\underline{z}$ is half way

for integer metric distances of practical interest in coding, e.g. Hamming distance or edit distance, the lower bound is "almost" achieved

$$\delta(\underline{x}, \underline{y}) = \left\lceil \frac{d(\underline{x}, \underline{y})}{2} \right\rceil$$

which seems to make the notion of distinguishability useless

but there might be holes in the string geometry

Distinguishabilities can be nasty even if $d(x, y)$ is nice (metric)

$d(x,y)$	a	b	c	d	e
a		1	1	1	$1/2$
b			$1/4$	$1/2$	1
c				$1/4$	$3/4$
d					$1/2$
e					

Distinguishabilities can be nasty even if $d(x, y)$ is nice (metric)

$d(x,y)$	a	b	c	d	e
a		1	1	1	$1/2$
b			$1/4$	$1/2$	1
c				$1/4$	$3/4$
d					$1/2$
e					

$\delta(x,y)$	a	b	c	d	e
a		1	$3/4$	$1/2$	$1/2$
b			$1/4$	$1/4$	$1/2$
c				$1/4$	$1/2$
d					$1/2$
e					

alas!

$$\delta(a, d) + \delta(d, b) = \tfrac{3}{4} < \delta(a, b) = 1$$

and now to an old distance

1967: Ž. Muljačić, Split-Spalato
Die Klassifikation der Romanischen Sprachen

$$d_n(\underline{x}, \underline{y}) = d_3(0, 0, \frac{1}{2} \ ; \ 0, 1, 1) = 0 + 1 + \frac{1}{2} = \frac{3}{2}$$

each linguistic feature can be either absent 0 , or present 1,
but some might be ill-defined,
intermediate logical values are needed, fuzziness is needed

x is the truth degree of "feature F is present in language L"
y is the truth degree of "feature F is present in language Λ", $x, y \in [0, 1]$

are x and y distinct?
is feature F present in one language and absent in the other?

$$d(x, y) = \Big[x \wedge (1 - y)\Big] \vee \Big[(1 - x) \wedge y\Big]$$

standard fuzzy choices, $\mathrm{OR} = \vee = \max$, $\mathrm{AND} = \wedge = \min$

x is the truth degree of "feature F is present in language L"
y is the truth degree of "feature F is present in language Λ", $x, y \in [0, 1]$

are x and y distinct?
is feature F present in one language and absent in the other?

$$d(x, y) = \Big[x \wedge (1 - y)\Big] \vee \Big[(1 - x) \wedge y\Big]$$

standard fuzzy choices, $\mathrm{OR} = \vee = \max$, $\mathrm{AND} = \wedge = \min$

$f(x) \doteq d(x, x) = x \wedge (1 - x)$ fuzziness of the logical value x
$f(x, y) \doteq f(x) \vee f(y) \leq \frac{1}{2}$ fuzziness of the couple x, y

$x \wedge y < \frac{1}{2} < x \vee y$ dissonant, else consonant

$n \geq 1, \quad \underline{x}, \underline{y} \in [0,1]^n$ n-length strings of logical values,
additive distances

$n \geq 1$, $\underline{x}, \underline{y} \in [0,1]^n$ n-length strings of logical values,
additive distances

Muljačić distance or fuzzy Hamming distance:

$$d(\underline{x}, \underline{y}) = d_n(\underline{x}, \underline{y}) = \sum_{i \in \mathcal{D}} \Big[1 - f(x_i, y_i)\Big] + \sum_{i \in \mathcal{C}} f(x_i, y_i)$$

$d(x, y)$ is a (fuzzy) *metric distance*

Muljačić distinguishability or fuzzy Hamming distinguishability:

$$\delta(\underline{x}, \underline{y}) = \frac{|\mathcal{D}|}{2} + f_{\mathcal{C}}(\underline{x}) \vee f_{\mathcal{C}}(\underline{y})$$

where $f_{\mathcal{C}}(\underline{x}) \doteq \sum_{i \in \mathcal{C}} f(x_i)$

$\delta(\underline{x}, \underline{y})$ is again a (fuzzy) *metric distance*

Muljačić distinguishability or fuzzy Hamming distinguishability:

$$\delta(\underline{x}, \underline{y}) = \frac{|\mathcal{D}|}{2} + f_{\mathcal{C}}(\underline{x}) \vee f_{\mathcal{C}}(\underline{y})$$

where $f_{\mathcal{C}}(\underline{x}) \doteq \sum_{i \in \mathcal{C}} f(x_i)$

$\delta(\underline{x}, \underline{y})$ is again a (fuzzy) *metric distance*

One always has a minimizing $\underline{z}$ over the ternary alphabet $\{0, \frac{1}{2}, 1\}$.
The computational effort to compute Muljačić distances
and distinguishabilities is only linear $\Theta(n)$

Why Muljačić distinguishabilities?

$\underline{x} = (1, 0)$, $\underline{y} = (\frac{1}{2}, \frac{1}{2})$, $\underline{u} = (1, 1)$
$d(\underline{x}, \underline{y}) = d(\underline{x}, \underline{u}) = 1$

the distance is the same

Why Muljačić distinguishabilities?

$\underline{x} = (1, 0)$, $\underline{y} = (\frac{1}{2}, \frac{1}{2})$, $\underline{u} = (1, 1)$
$d(\underline{x}, \underline{y}) = d(\underline{x}, \underline{u}) = 1$

the distance is the same

That's why!

$\delta(\underline{x}, \underline{y}) = 1 \neq \delta(\underline{x}, \underline{u}) = \frac{1}{2}$

distinguishabilities are different

let's change, tentatively

"abstract" conjunctions $\wedge$ and disjunctions $\vee$
$\top$-norms $\top$ and $\top$-conorms $\perp$, same negation

commutative and associative, monotone non-decreasing
$1 \top x = 0 \perp x = x$, $0 \top x = 0$, $1 \perp x = 1$
duality: we impose De Morgan laws

what happens of $d(x, y) = (x \top \overline{y}) \perp (\overline{x} \top y)$?

nothing on the crisp border of the unit square, $d(x, y) = |x - y|$

Łukasiewicz: $x \perp y = x + y$ for $x + y \leq 1$, else 1
$f(x) \doteq d(x, y) = 0$, $d(x, y) = |x - y|$
too good to be true, i.e. too crisp to be fuzzy

probabilistic: $x \top y = xy$
$f(x) = x^2$, possibly $d(x, x) < d(x, y)$
good fuzziness but a metric flop, one loses bounds on $\delta(x, y)$

drastic: $x \top y = 0$ unless on the border
$f(x) = 0$, possibly $d(x, z) + d(z, y) < d(x, y)$
bad fuzziness and a metric disaster

nil-potent minimum: $x \top y = x \wedge y$ for $x + y \geq 1$, else 0
one re-obtains Muljačić distance

Łukasiewicz: $x \perp y = x + y$ for $x + y \leq 1$, else 1
$f(x) \dot{=} d(x, y) = 0$, $d(x, y) = |x - y|$
too good to be true, i.e. too crisp to be fuzzy

probabilistic: $x \top y = xy$
$f(x) = x^2$, possibly $d(x, x) < d(x, y)$
good fuzziness but a metric flop, one loses bounds on $\delta(x, y)$

drastic: $x \top y = 0$ unless on the border
$f(x) = 0$, possibly $d(x, z) + d(z, y) < d(x, y)$
bad fuzziness and a metric disaster

nil-potent minimum: $x \top y = x \wedge y$ for $x + y \geq 1$, else 0
one re-obtains Muljačić distance

So:

too good, too bad, much too bad, déja vu

A general $\top$-norm:
given a region $\mathcal{R}$ such that one can exchange and/or diminish x and y, ...

one sets $x \top y = 0$ in $\mathcal{R}$ and $x \top y = x \wedge y = \min(x, y)$ outside $\mathcal{R}$

A general $\top$-norm:
given a region $\mathcal{R}$ such that one can exchange and/or diminish x and y, ...

one sets $x \top y = 0$ in $\mathcal{R}$ and $x \top y = x \wedge y = \min(x, y)$ outside $\mathcal{R}$

The R-minimum norms give back Muljačić distance whenever $\mathcal{R} \cap \neg\mathcal{R} = \emptyset$

The triangle inequality falls whenever $|\mathcal{R} \cap \neg\mathcal{R}| \geq 2$

in the (uninteresting) case $|\mathcal{R} \cap \neg\mathcal{R}| = 1$, i.e. $\mathcal{R} \cap \neg\mathcal{R} = (\frac{1}{2}, \frac{1}{2})$, ...

Remarkable *déja vu* cases:

Timișoara norms

$\mathcal{R}$ made up by couples for which $x \vee y < \alpha \leq \frac{1}{2}$

Generalized nilpotent minimum

$\mathcal{R}$ made up by couples for which $x + y < \beta \leq \frac{1}{2}$

All of this enhances the relevance of Muljačić distances!

Remarkable *déja vu* cases:

Timişoara norms

$\mathcal{R}$ made up by couples for which $x \vee y < \alpha \leq \frac{1}{2}$

Generalized nilpotent minimum

$\mathcal{R}$ made up by couples for which $x + y < \beta \leq \frac{1}{2}$

All of this enhances the relevance of Muljačić distances!

By the way:
$d_{\mathrm{Mul}}(\underline{x}, \underline{y}) = d_{\mathrm{taxi}}(\underline{x}, \underline{y}) + \sum_i f(x_i) \wedge f(y_i)$

The main contribution is very technical and possibly rather boring.

In short:

we have considered alternative logical operators, T-norms and T-conorms of MV logics, including the most popular ones. Results are either untenable (faulty distances) or lead to the very same fuzzy distance as in Muljačić case.

And now at last language evolution and fuzzy classification:

Powerful metaphor? noisy channel $\approx$ linguistic evolution in time

Forwards: what will happen in the future? One needs distances, more generally dissimilarities

Backwards: what has happened in the past, can one decode to ancestors: one needs also distinguishabilities

And now at last language evolution and fuzzy classification:

Powerful metaphor? noisy channel $\approx$ linguistic evolution in time

Forwards: what will happen in the future? One needs distances, more generally dissimilarities

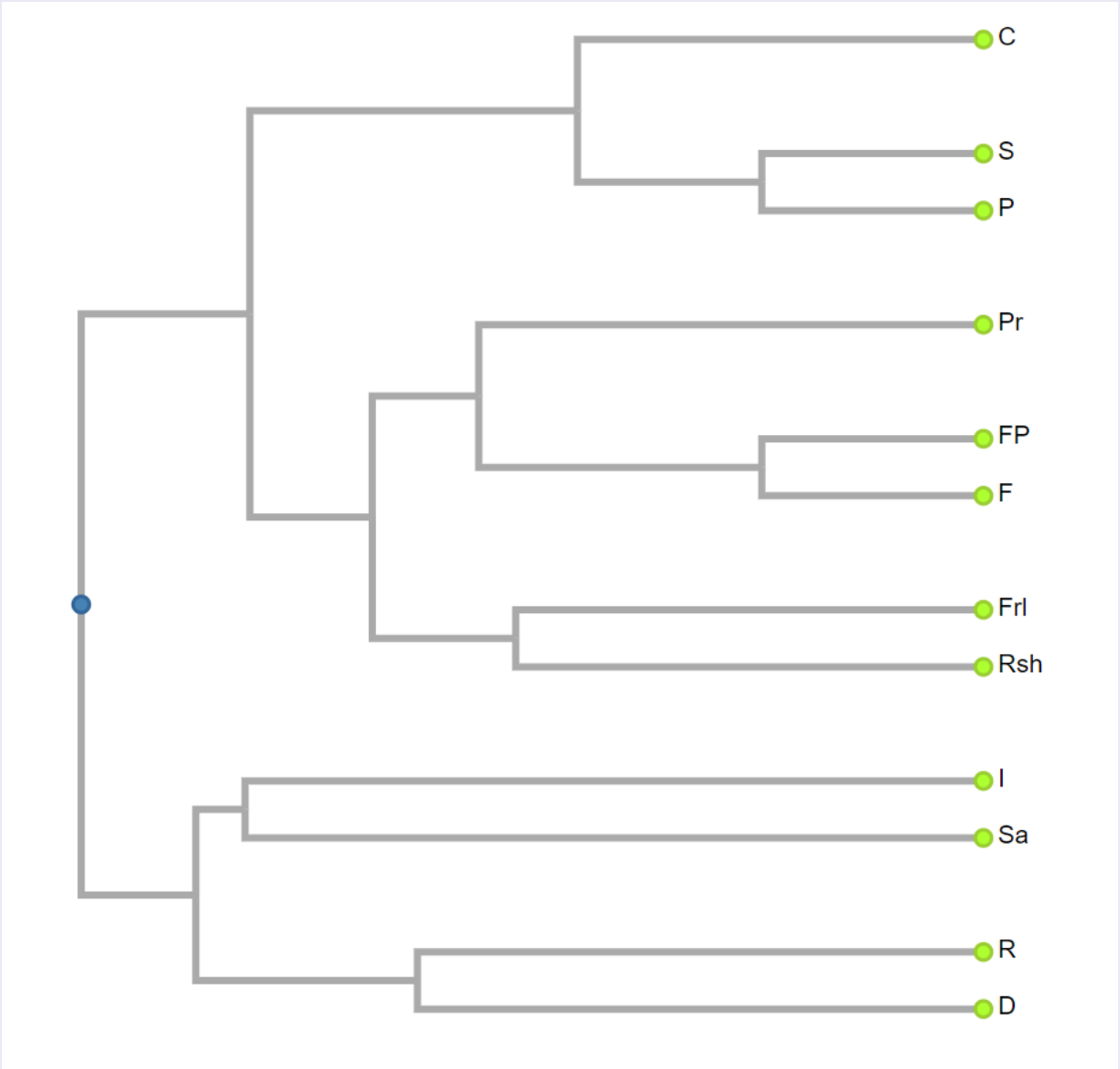
Backwards: what has happened in the past, can one decode to ancestors: one needs also distinguishabilities

If no fuzziness is to be dealt with, as usual in linguistic and bioinformatics, distinguishabilities are forgotten to no harm, but...

two matrices

$2d/4\delta$	R	D	I	Sa	Frl	Rsh	Pr	FP	F	C	S	P
R												
D	23/24											
I	37/38	26/26										
Sa	33/34	32/32	30/30									
Frl	37/38	26/26	32/32	38/38								
Rsh	38/40	25/26	31/32	40/42	19/20							
Pr	38/42	27/30	33/36	41/44	21/24	24/26						
FP	47/50	32/34	36/38	52/54	22/24	23/24	20/22					
F	54/56	39/40	43/44	59/60	29/30	30/30	21/24	9/12				
C	37/38	26/26	32/32	38/38	22/22	23/24	19/22	22/24	29/30			
S	31/32	30/30	30/30	34/34	30/30	31/32	26/30	34/36	41/42	14/14		
P	34/34	39/40	33/34	41/42	39/40	32/34	30/34	31/34	38/40	19/20	9/10	

classification



decoding to ancestors

Let us ask two (only jocular) questions, just to use our tools:

Is Dalmatic a dialect of Romanian or of Italian? The best (jocular) answer is Romanian, however unreliable, because the distinguishability between Romanian and Italian is smaller that their distances from Dalmatic

Is Franco-provençal a dialect of French or of Italian? The best (jocular) answer is French, this time reliable, at least within our (linguistically untenable) assumptions

It would be nice to have strings for old languages, say Anglosaxon, Chaucer's English and modern English

FUZZ-IEEE 2017, applications to error-correction codes

Synasc 2017, accepted: language classification with fuzziness and *irrelevance*, Jaccard-like distances

Ongoing experiments on real, up-to-date and extensive data, Human Language Technologies Research Center, Bucharest

Merci, grazie,

mulțumesc, thanks !