

Automatic measurement criteria

- Measurement with respect to generated “truth”.
- **TER** token error rate:
the percentage of original unsplit tokens whose expansion to words does not completely match the expansion to words in truth.
- **WER** word error rate:
the percentage of wrong words in an expansion (including insertions, deletions and substitutions) with respect to truth.

Baseline results

LDC tools : LDC text conditioning tools

Festival : 1.4.0 released text analyser

	LDC tools		Festival	
	TER	WER	TER	WER
nantc	–	2.88	1.00	1.38
classifieds	–	30.81	30.09	33.48
pc110	–	22.36	14.37	32.62
rfr	–	9.06	6.28	16.19

Domain dependent model

- domain independent splitter
- CART tag classifier with letter language model features
- EXPNs by WFST
- Language model

	festival		m4	
	TER	WER	TER	WER
nantc	1.00	1.38	0.39	0.82
classifieds	30.09	33.48	7.00	9.71
pc110	14.37	32.62	3.66	9.25
rfr	6.28	16.19	0.94	2.07

Removing components

m4.nolm: no language model (most prob EXPN)

m4.noef: no letter language models feats

m4.noeflm: no LM and no LLM feats

	m4		m4.nolm		m4.noef		m4.noeflm	
	TER	WER	TER	WER	TER	WER	TER	WER
nantc	0.39	0.82	0.39	0.81	0.38	0.78	0.38	0.78
classifieds	7.00	9.71	6.82	9.70	7.55	10.39	7.41	10.42
pc110	3.66	9.25	3.63	9.25	3.93	10.90	3.90	10.90
rfr	0.94	2.07	0.93	2.06	0.88	2.07	0.88	2.07

Giving truth

m4.nosplt: uses hand labeled splits

m4.nost: uses hand labeled splits and actual tags

	m4		m4.nosplt		m4.nost	
	TER	WER	TER	WER	TER	WER
nantc	0.39	0.82	0.20	0.44	0.03	0.06
classifieds	7.00	9.71	5.40	6.35	3.15	4.24
pc110	3.66	9.25	2.58	4.61	0.49	0.75
rfr	0.94	2.07	0.59	1.11	0.16	0.24

Cross-domain models

m4.domin: nantc models

m4.dominE: nantc models with domain EXPNs

	festival		m4		m4.domin		m4.dominE	
	TER	WER	TER	WER	TER	WER	TER	WER
nantc	1.00	1.38	0.39	0.82	0.39	0.82	0.39	0.82
classifieds	30.09	33.48	7.00	9.71	25.20	29.11	19.69	21.18
pc110	14.37	32.62	3.66	9.25	12.35	18.69	12.09	18.07
rfr	6.28	16.19	0.94	2.07	2.71	4.66	2.32	4.14

Unsupervised domain models

Building models from unlabeled data

- Label tokens with nante CART tag classifier
- Relabel alphabets with best LLM prediction
- Build EXPN expander from plain text and labeled EXPNs
- Build words with best EXPN expansion
- Build LM from full expanded words
- Run with multiple EXPNs and LM to choose

	TER	WER
m4	7.00	9.71
us1.lm	12.50	13.40
us1.nolm	12.64	13.50
us2.EXPNIlist	10.58	13.51
m4.dominE	19.69	21.18

Results

- Marked up databases
- Tools to help label databases
- Tools and methods for building models
- 4 domain models
- Text expander better than LDC or Festival
- Tools and methods for building unsupervised models

And finally ...

Pre-WS99: Festival-based ads expander

- Searching for appropriate properties. Found 1 matching
57 ST E/1st & 2nd Ave Huge
drmm 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg
\$187K, maint \$868, utils
incl. Call Bkr Peter 914-428-9054.
- Do you wish more details?
- Or will you call back later?

festival expansion

57 ST E/1st & 2nd Ave Huge

drmn 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg

\$187K, maint \$868, utils

incl. Call Bkr Peter 914-428-9054.

fifty seven saint E. slash frst & second Avenue Huge

d. r. m. n. one B. R. seven five zero +. s. f. lots of sun &

e. l. s. t. s. Sundeck & Indry facils Askg

\$. one eight seven K. maint eight hundred sixty

eight dollars utils

incl. Call B. k. r. Peter nine one four four two eight

nine zero five four

Later ...

Post-WIS99: NSW-based ads expander (m4)

- Searching for appropriate properties. Found 1 matching
57 ST E/1st & 2nd Ave Huge
drmn 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg
\$187K, maint \$868, utils
incl. Call Bkr Peter 914-428-9054.
- This property is still for sale. Do you wish more details?

m4 expansion

57 ST E/1st & 2nd Ave Huge
drmm 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg
\$187K, maint \$868, utils
incl. Call Bkr Peter 914-428-9054.

fifty seven STREET E / first AND second AVENUE Huge
D R M N one BEDROOM seven fifty PLUS SQUARE FOOT,
lots of sun AND CLOSETS. Sundeck AND LAUNDRY FA-
CILITIES. ASKING

one hundred eighty seven thousand dollars, MAINTENANCE
eight hundred sixty eight dollars, UTILITIES INCLUDED. Call
BROKER Peter nine one four, four two eight,
nine zero five four.

A little later ...

Post-WS99: NSW-based unsupervised ads expander

- Searching for appropriate properties. Found 1 matching
57 ST E/1st & 2nd Ave Huge
drmn 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg
\$187K, maint \$868, utils
incl. Call Bkr Peter 914-428-9054.

us1 expansion

57 ST E/1st & 2nd Ave Huge
drmn 1 BR 750+ sf, lots of sun &
clsts. Sundeck & Indry facils. Askg
\$187K, maint \$868, utils
incl. Call Bkr Peter 914-428-9054.

fifty seven STREET END / first AND second AVENUE Huge
DOORMAN one BEDROOM seven hundred fifty + SQUARE
FOOT ,

lots of sun AND CLOSETS. Sundeck AND LAUNDRY facils.
ASKING

one hundred eighty seven thousand dollars , MAINTAINED
eight hundred sixty eight dollars , UTILITIES
INCLUDED. Call BAKER Peter nine one four, four two eight,
nine zero five four.