Hands on with the Switchboard Corpus

Ye Tian, Université Paris Diderot

Day 1 PM

The Switchboard Corpus

- The Switchboard-1 Telephone Speech Corpus (LDC97S62) was originally collected by Texas Instruments in 1990.
- A collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States.
- A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished.
- About 70 topics were provided(1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.

SWB dialog act corpus

- The Switchboard Dialog Act Corpus (SwDA) extends the Switchboard-1 Telephone Speech Corpus, Release 2, with turn/utterance-level dialog-act tags.
- The dialogue act (speech act type of each utterance) tagging allow the modelling of shallow discourse structure.
- The corpus is also annotated for disfluency and nonverbal elements such as laughter.

Getting the corpus

- A good tutorial:
- http://compprag.christopherpotts.net/swda.html

- The SDA trascripts are a free download: http://www.stanford.edu/~jurafsky/swb1_dialogact_annot.tar.gz
- The recommended method by Jurafsky and Potts is to use python NLTK package (Natural Language Toolkit). For instructions and information, follow http://compprag.christopherpotts.net/swda.html and http://www.nltk.org/ .
- In this tutorial we use R.
  - I have done some minor processing on the original data file
  - Download act tag list.csv and alltrans_data.csv. The rest can be obtained from the link above.

Let's look at a data file

- Without dialogue tags
- sw2001A-ms98-a-trans
- sw2001A-ms98-a-word

- With dialogue act tags and other annotations:
- sw_0001_4325.utt

- metadata

Annotations – dialogue acts

- See coder's manual for detailed informations

http://web.stanford.edu/~jurafsky/ws97/manual.august1.html

List of top dialogue act tags

Speech act	SWBD	Example	Cnt	%
Statement-non-opinion	sd	Me, I'm in the legal department.	72,824	36%
Acknowledge (Backchannel)	b	Uh-huh.	37,096	19%
Statement-opinion	sv	I think it's great	25,197	13%
Agree/Accept	aa	That's exactly it.	10,820	5%
Abandoned or Turn-Exit	% -	So, -	10,569	5%
Appreciation	ba	I can imagine.	4,633	2%
Yes-No-Question	qy	Do you have to have any special training?	4,624	2%
Non-verbal	x	[Laughter], [Throat_clearing]	3,548	2%
Yes answers	ny	Yes.	2,934	1%
Conventional-closing	fc	Well, it's been nice talking to you.	2,486	1%
Wh-Question	qw	Well, how old are you?	1,911	1%
No answers	nn	No.	1,340	1%
Response Acknowledgement	bk	Oh, okay.	1,277	1%
Hedge	h	I don't know if I'm making any sense or not.	1,182	1%
Declarative Yes-No-Question	qy^d	So you can afford to get a house?	1,174	1%

Annotations – PoS tags

- The Penn Treebank Part-of-Speech Tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

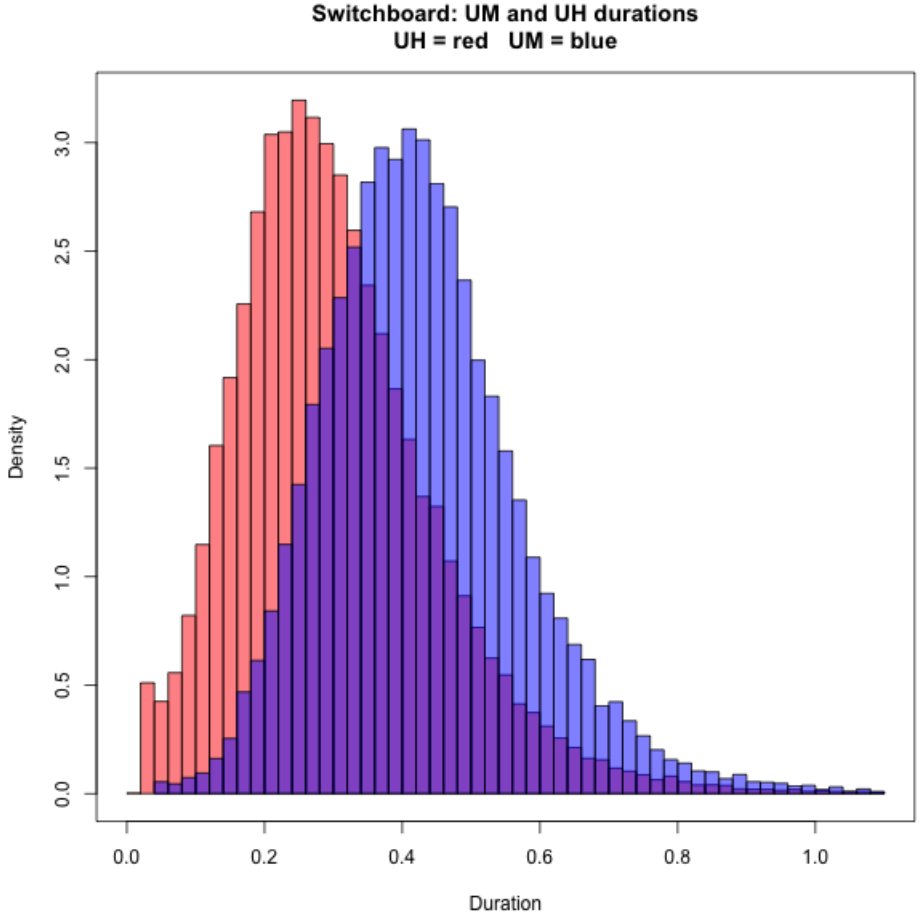
Figure 9.1 Penn Treebank part-of-speech tags (including punctuation).

Annotations – parsed trees

```
>>> [tree.pprint() for tree in utt.trees]
["(S
  (CC And)
  (NP-SBJ (PRP it))
  (VP
    (BES 's)
    (NP-PRD
      (NP (DT a) (JJ small) (NN office))
      (SBAR
        (WHNP-1 (WDT that))
        (S
          (NP-SBJ (PRP she))
          (VP (VBZ works) (PP-LOC (RB in) (NP (-NONE- *T*-1))))))))
  (-DFL- E_S))"]
>>> utt.tree_lemmas(wn_lemmatize=True)
[('And', 'CC'), ('it', 'PRP'), ("'s", 'BES'), ('a', 'DT'), ('small', 'JJ'), \
('office', 'NN'), ('that', 'WDT'), ('she', 'PRP'), ('works', 'VBZ'), ('in', 'RB'), \
('*T*-1', '-NONE-'), ('E_S', '-DFL-')]
```

Example studies

- http://languagelog.ldc.upenn.edu/nll/?p=14991



Let's do something

- First a bit of regular expressions
- Search for a word/phrase
- Search for dialogue acts
- Search for words with specific PoS

- Files used:
- Switchboard search engine.R
- swda tagged_treebank search engine.R

Regular expressions

- Regular expressions are the standard notation for characterizing text sequences.
- First developed by Kleene (1956)
- We use regular expressions to search for patterns in a corpus.
- The simplest kind of regular expression is a sequence of characters. E.g. /linguistics/.
- Regular expressions are case sensitive.
- Let's work on some examples
- https://en.wikibooks.org/wiki/R_Programming/Text_Processing

References

- Kleene, Stephen C. (1956). Shannon, Claude E.; McCarthy, John, eds. Representation of Events in Nerve Nets and Finite Automata. Automata Studies. Princeton University Press. pp. 3–42.