What is domain?

Rebecca Dridan

DELPH-IN summit 2014, Tomar

The WeSearch project

Focus: large-scale parsing of user-generated content

Relevant aims:

- Explore and document factors often ignored: accuracy vs. efficiency tradeoffs; domain effects; etc.
- Develop ~~domain~~ parser adaptation for user-generated content.

So far:

- ubertagging to speed up parsing
- creation of WeSearch Data Collection (WDC) which distinguishes genre: wiki and blog, and domain: Linux and NLP

Domain Adaptation

- Gildea (2001): statistical parsers trained on WSJ data are less accurate test on non-WSJ data.
- Almost all work since has compared WSJ or Genia (large cohesive domains) except Foster et. al.
- *Domain* is generally defined as 'section of a corpus'.

Note: Our parser (PET+ERG) is not a purely statistical parser.

Corpus comparison

from Comparing Corpora. Adam Kilgarrif.
International Journal of Corpus Linguistics, 6:1 (2001), 97–133.

- Corpora are compared by their word use, since anything else requires annotation.
- Qualitative comparison: which words are most representative of each corpus.
  - needs methods that consider distribution and frequency.
  - should not assume null hypothesis that words are randomly distributed.
- Quantatative comparison: which corpora are most similiar/most different to each other.
  - requires looking at both intra and inter corpus similarities.
  - showed χ^2 was the best in a known similiarity experiment.

Average χ^2 over top 500 words

	WLB	**WLW**	**WNB**	**WNW**
WLB	333	344	296	563
WLW		113	149	235
WNB			124	124
WNW				145

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Average χ^2 over top 500 words

	WLB	WLW	WNB	WNW
WLB	333	344	296	563
WLW		113	149	235
WNB			124	124
WNW				145

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Average χ^2 over top 500 words

	WLB	WLW	WNB	WNW
WLB	333	344	296	563
WLW		113	149	235
WNB			124	124
WNW				145

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Average χ^2 over top 500 words

	WLB	WLW	WNB	WNW
WLB	333	344	296	563
WLW		113	149	235
WNB			124	124
WNW				145

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Average χ^2 over top 500 words

	WLB	**WLW**	**WNB**	**WNW**	**WSJ**
WLB	333	344	296	563	554
WLW		113	149	235	294
WNB			124	124	300
WNW				145	322
WSJ					190

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Average χ^2 over top 500 words

	WLB	WLW	WNB	WNW	WSJ
WLB	333	344	296	563	554
WLW		113	149	235	294
WNB			124	124	300
WNW				145	322
WSJ					190

	Blog	Wiki	WSJ
Blog	263	295	427
Wiki		184	308
WSJ			190

	NLP	Linux	WSJ
NLP	129	310	311
Linux		285	424
WSJ			190

$$\chi^2 = \sum_{i \in \text{words}} \frac{(O_i - E_i)^2}{E_i}$$

Most salient words using log-likelihood

NLP vs. Linux		Wiki vs. Blog	
I	9575	you	17759
Ubuntu	8076	I	15937
sudo	5451	your	7294
install	4763	sudo	5331
file	3640	we	5229
Linux	3551	Ubuntu	5166
that	2879	n't	4556
n't	2809	of	3691
command	2760	install	3111
Windows	2630	my	2906

$$\text{log-likelihood} = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

Most salient words using log-likelihood

NLP vs. Linux		in Wiki	
I	9575	language	1190
Ubuntu	8076	Esperanto	894
sudo	5451	English	832
install	4763	words	823
file	3640	languages	617
Linux	3551	algorithms	600
that	2879	German	570
n't	2809	Windows	517
command	2760	linguistics	493
Windows	2630	Google	490

$$\text{log-likelihood} = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

Most salient words using log-likelihood

Wiki vs. Blog		in NLP	
you	17759	I	4445
I	15937	software	1511
your	7294	n't	1456
sudo	5331	you	1430
we	5229	we	1190
Ubuntu	5166	German	1023
n't	4556	's	1017
of	3691	Esperanto	899
install	3111	my	800
my	2906	Linux	791

$$\text{log-likelihood} = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

Syntax-based comparison using WDC treebank

Profile	Items	Len.	TTR	OOV
WLB_d	452	10.8	0.34	11.9
WLW_d	431	18.1	0.28	9.2
WNB_d	474	13.7	0.31	6.4
WNW_d	459	16.1	0.28	6.6

TTR: type:token ratio
OOV: out of vocabulary tokens (not in lexicon)

Syntax-based comparison using WDC treebank

Profile	Items	Len.	TTR	OOV
WLB_d	452	10.8	0.34	11.9
WLW_d	431	18.1	0.28	9.2
WNB_d	474	13.7	0.31	6.4
WNW_d	459	16.1	0.28	6.6
WSJ_{20}	1721	19.8	0.20	7.2

TTR: type:token ratio
OOV: out of vocabulary tokens (not in lexicon)

Most salient labels using log-likelihood

NLP vs. Linux

n_-_pn-gen_le	hdn_bnp-pn_c
n_-_pn_le	n_sg_i_ilr
cm_np-vp_that_le	np_hdn_nme_cpd_c
	np_hdn_cpd_c

Wiki vs. Blog

n_-_pr_i_le	hdn_bnp-qnt_c
n_-_pr_you_le	v_n3s-bse_ilr
n_-_pr_we_le	w_bang_plr
d_-_poss_your	

Corpus comparison summary

- The Linux blog data is different
  - to itself and
  - to the rest of the data.
- This makes it difficult to say anything meaningful about blog vs. wiki or Linux vs. NLP (ie genre or domain differences).
- The main differences visible seem to be in pronoun use (blogs), proper names (Linux), and possibly noun compounds (Linux).

Parse Accuracy: Parseval

	Redwoods		Deepbank	
Profile	**F1**	**SA**	**F1**	**SA**
WLB_d	81.2	35.6	81.8	29.2
WNB_d	84.3	40.3	83.3	39.2
WLW_d	85.0	35.5	83.8	31.8
WNW_d	86.7	45.1	84.1	37.5
WSJ_{20}	81.1	24.4	89.5	45.1

Error analysis by text type (semantic argument)

Proportion of errors (%)

WLB

udef_q	15.1
prep ARG1	8.6
compound ARG2	7.9
proper_q	8.6
verb ARG1	8.6

WLW

udef_q	18.2
prep ARG1	9.0
compound ARG2	8.5
compound ARG1	7.4
verb ARG2	6.8

WNB

udef_q	15.5
prep ARG1	8.7
verb ARG2	6.5
adj ARG1	4.6
verb ARG1	4.4

WNW

udef_q	18.8
prep ARG1	10.5
compound ARG2	7.2
compound ARG1	6.1
verb ARG2	5.4

Error analysis by text type (semantic argument)

Accuracy of arguments (%)

WLB

udef_q	15.1	75.4
prep ARG1	8.6	68.8
compound ARG2	7.9	68.4
proper_q	8.6	79.8
verb ARG1	8.6	87.1

WLW

udef_q	18.2	82.7
prep ARG1	9.0	80.8
compound ARG2	8.5	78.7
compound ARG1	7.4	81.3
verb ARG2	6.8	90.4

WNB

udef_q	15.5	82.7
prep ARG1	8.7	80.3
verb ARG2	6.5	89.2
adj ARG1	4.6	90.6
verb ARG1	4.4	92.7

WNW

udef_q	18.8	83.6
prep ARG1	10.5	80.2
compound ARG2	7.2	80.0
compound ARG1	6.1	83.5
verb ARG2	5.4	93.3

Conclusions

- It is difficult to draw any conclusions about genre versus domain *with the data we have*.
- Parse accuracy on WDC is lower than for WSJ data with in-domain training data.
- There are slight trends between text types, but mostly variance overwhelms differences.
- Hard to determine what 'in-domain' data is for a text type as heterogeneous as Linux blog data.
- Improving PP attachment accuracy will improve parse accuracy overall, but
  - it is difficult
  - it doesn't appear to be affected by domain/genre differences, at least with our parse selection features.