The effect of topic on documents in the Cambridge Learner Corpus

Andrew Caines & Paula Buttery; apc38@cam.ac.uk, pjb48@cam.ac.uk

Computational Linguistics Cluster, Department of Theoretical & Applied Linguistics
Institute for Research in Automated Language Teaching and Assessment (ALTA)

OUR RESEARCH QUESTIONS:

- *Given our set of 4 labels, does document topic affect the distribution of lexico-syntactic features in learner corpora?*
- *Can we train a naive Bayes (NB) classifier to label unseen documents accurately, based on lexico-syntactic features?*
- *Should topic be controlled for in learner corpus studies?*

THE DATA:

- CLC-2009-B1, a subset of the Cambridge Learner Corpus;
- Cambridge English exam scripts from the year 2009;
- CEFR level B1 (PET, PETfS, BECP, SfLe3);
- 3427 documents, 288k words;
- each answer matched with its exam question → topic label;
- mean sentence length = 9.5 words (*cf.* A2=6.7, B2=12.3)
- mean document length = 84 words (*cf.* A2=40, B2=121)

TOPIC TAXONOMY:

- **commerce** – business, administration, sales and marketing (*i.e.* BECP examination scripts);
- **narrative** – creative story writing, often starting with a set sentence (*e.g.* It was getting dark and I was completely lost);
- **personal** – requires the candidate to relate autobiographical events, to role play in such events, or to give subjective views of cultural objects such as films, restaurants or works of literature;
- **society** – relates to wider issues such as the education system, public transport or the environment.

VERB SUBCATEGORIZATION FRAMES:

- SCFs distinguish verb argumentation patterns, thereby encoding constructions;
- set of 163 designed by Ted Briscoe and John Carroll, widely used in experimental work;
- *e.g.*
  - Stephen surfs; Frame 22, INTRANS
  - Andrew bought a juicer; Frame 24, NP
  - Lindsay put Harvey on the floor; Frame 49, NP-PP
- we extracted SCFs from CLC-2009-B1 using the RASP System, and paired each verb with one or more SCFs;
- *e.g.* surf_22, buy_24, put_49
- due to ambiguity, some verbs associated with >1 SCF;
- write_56v49
  - SCF 56 = NP-TO-NP, *he wrote a letter to her*
  - SCF 49 = NP-PP, *he was writing his letter to the last possible moment*
- hope_33v32
  - SCF 33 = NP-INF-OC, *she hoped to run the race*
  - SCF 32 = NP-INF, *she hoped to run later on.*

CLASSIFYING BY LEXICAL FEATURES:

label	precision	recall	F-measure
commerce	0.9512	0.9499	0.9492
narrative	0.4595	0.9928	0.6244
personal	1.0	0.7836	0.8700
society	0.2164	1.0	0.3398
		overall accuracy = 0.8107	

Figure : Precision, recall and F-measures for each topic label, from a naive Bayes classifier trained on lexical features in CLC-2009-B1, using 10-fold cross-validation

CLASSIFYING BY SYNTACTIC FEATURES:

label	precision	recall	F-measure
commerce	0.8756	0.6510	0.7463
narrative	0.4256	0.8915	0.5657
personal	0.9842	0.6715	0.7907
society	0.0880	0.9374	0.1589
		overall accuracy = 0.6851	

Figure : Precision, recall and F-measures for each topic label, from a naive Bayes classifier trained on SCFs in CLC-2009-B1, using 10-fold cross-validation

'HIGH INFO' FEATURES:

word	discriminates	strength
learn_22	society:personal	98:1
educate_24	society:personal	98:1
learn_24v51	society:personal	88:1
get_87v96	narrative:personal	84:1
fall_95	narrative:personal	81:1
teach_24v51	society:personal	70:1
think_153	society:personal	70:1
develop_22	society:personal	70:1
study_33v32	society:personal	70:1
catch_24v51	narrative:personal	68:1
prefer_22	commerce:personal	62:1

Figure : Selection of highly informative features from the naive Bayes classifiers trained on lexical features and SCFs in CLC-2009-B1

CONCLUSIONS:

i, we identified a set of 4 topic labels in CLC-2009-B1 and trained an NB classifier on labelled lexical features (words) to a high degree of accuracy (81.1%);

ii, we also analysed verb argumentation patterns and trained an NB classifier on verb-SCF pairings to a reasonable degree of accuracy (68.5%);

iii, analysis of the highly informative features indicates the underlying lexico-syntactic differences;

iv, topic needs to be controlled for in learner corpus research, at least for investigations of lexis and verb argumentation patterns.

generously funded by Cambridge English Language Assessment