Appendix A

Data Sets

A number of data sets were used for experimental evaluation purposes in this thesis. The section summarises the data contained in each.

A.1 Language modelling

A.1.1 Enron e-mail corpus

This dataset consists of approximately 500,000 e-mail messages sent by approximately 150 members of the staff of the Enron corporation, and was originally released as part of a legal investigation of the company. Preprocessing has been done to remove attachments [43].

A.1.2 Canterbury corpus

The Canterbury corpus [104] [1] aims to provide a standard data set for comparison of compression algorithms. The corpus consists of 11 files of varying time, summarised in Table A.1. Analysis in this thesis concentrates on the file `alice29.txt` as an example of natural English.

File	Abbreviation	Category	Size
alice29.txt	text	English text	152089
asyoulik.txt	play	Shakespeare	125179
cp.html	html	HTML source	24603
fields.c	Csrc	C source	11150
grammar.lsp	list	LISP source	3721
kennedy.xls	Excl	Excel Spreadsheet	1029744
lcet10.txt	tech	Technical writing	426754
plrabn12.txt	poem	Poetry	481861
ptt5	fax	CCITT test set	513216
sum	SPRC	SPARC Executable	38240
xargs.1	man	GNU manual page	4227

Table A.1: Files making up the Canterbury corpus.

A.2 Information retrieval

A.2.1 TREC corpus

The TREC corpus is a large data set consisting of articles taken from a variety of newswire and other sources. The data set is the basis of the TREC information retrieval competition [113]. The data is supplied on five CD-ROMs, although only discs 4 and 5 are used in the evaluations presented above. This data set consists of 528,155 documents spanning a total of 165,363,765 terms from a vocabulary of size 629,469 (after stop words removal).

Also provided are a number of query strings consisting of three parts, a title, description and narrative. Ground truth judgements are available concerning whether or not each of the documents is relevant to each of the queries. Two sets of fifty queries (referred to as TREC-7 and TREC-8) were used in this thesis.

A.2.2 Cranfield corpus

The Cranfield corpus is a relatively small information retrieval corpus consisting of 1400 abstracts on aeronautical engineering topics. The documents contain a total of 136935 terms from a vocabulary of size 4,632 (after stop word removal).

The Cranfield corpus also contains a set of 225 query strings with ground truth relevance judgements. Relevances are assigned on four levels, but for purposes of comparison to the TREC data set this was converted to a binary value for the experiments presented in this thesis.

Appendix B

Stop Word List

The following list of 319 stop words was used in the information retrieval experiments. It was made available by the Information Retrieval Group at the University of Glasgow [103].

a	anyone	both	eleven	four	however
about	anything	bottom	else	from	hundred
above	anyway	but	elsewhere	front	i
across	anywhere	by	empty	full	ie
after	are	call	enough	further	if
afterwards	around	can	etc	get	in
again	as	cannot	even	give	inc
against	at	cant	ever	go	indeed
all	back	co	every	had	interest
almost	be	computer	everyone	has	into
alone	became	con	everything	hasnt	is
along	because	could	everywhere	have	it
already	become	couldnt	except	he	its
also	becomes	cry	few	hence	itself
although	becoming	de	fifteen	her	keep
always	been	describe	fify	here	last
am	before	detail	fill	hereafter	latter
among	beforehand	do	find	hereby	latterly
amongst	behind	done	fire	herein	least
amoungst	being	down	first	hereupon	less
amount	below	due	five	hers	ltd
an	beside	during	for	herself	made
and	besides	each	former	him	many
another	between	eg	formerly	himself	may
any	beyond	eight	forty	his	me
anyhow	bill	either	found	how	meanwhile

might
mill
mine
more
moreover
most
mostly
move
much
must
my
myself
name
namely
neither
never
nevertheless
next
nine
no
nobody
none
noone
nor
not
nothing
now
nowhere
of
off
often
on
once
one
only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own
part
per
perhaps
please
put
rather
re
same
see
seem
seemed
seeming
seems
serious
several
she
should
show
side
since
sincere
six
sixty
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
such
system
take
ten
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
thick
thin
third
this
those
though
three
through
throughout
thru
thus
to
together too
top
toward
towards
twelve
twenty
two
un
under
until
up
upon
us
very
via
was
we
well
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
would
yet
you
your
yours
yourself
yourselves

Bibliography

Articles

[1] Ross Arnold and Tim Bell. A corpus for the evaluation of lossless compression algorithms. Technical report, Department of Computer Science, University of Canterbury, 1997.

[2] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1001–1008, 1983.

[3] Adrian Barbu and Song-Chun Zhu. Graph partition by Swendsen–Wang cuts. In *Proceedings of the International Conference on Computer Vision*, 2003.

[4] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584, 2002.

[5] Timothy Bell, I. H. Witten, and J. G. Cleary. Modeling for text compression. *ACM Computing Surveys*, 21(4):557–592, 1989.

[6] Timothy C. Bell, John G. Cleary, and Ian H. Witten. *Text Compression*. Prentice Hall, 1990.

[7] Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. A neural probabilistic language model. In *Advances in Neurological Information Processing Systems*, volume 13, pages 932–938, 2000.

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[9] J. Bentley, D. Sleator, R. Tarjan, and V. Wei. A locally adaptive data compression scheme. *Communications of the ACM*, 29:320–330, 1986.

[10] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *1999 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999.

[11] David Blackwell and James B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973.

[12] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. In *Advances in Neurological Information Processing Systems*, volume 14, 2002.

[13] John Blitzer, Amir Globerson, and Fernando Pereira. Distributed latent variable models of lexical co-occurences. In *Proceedings of the Tenth International Workshop on Artifical Intelligence and Statistics*, 2005.

[14] Sregey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Combuter Networks and ISDN Systems*, 30:107–117, 1998.

[15] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[16] James P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual Internation ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[17] Stanley F. Chen. *Building Probabilistic Models For Natural Language*. PhD thesis, Harvard University, 1996.

[18] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modelling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[19] Kenneth W. Church and William A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54, 1991.

[20] J. G. Cleary and W. J. Teahan. Unbounded length contexts for PPM. *The Computer Journal*, 40:67–75, 1997.

[21] John G. Cleary and Ian H. Witten. A comparison of enumerative and adaptive codes. *IEEE Transactions on Information Theory*, 30(2):306–315, 1984.

[22] Philip J. Cowans. Information retrieval using hierarchical Dirichlet processes. In Kalervo Jarvelin, James Allan, Peter Bruza, and Mark Sanderson, editors, *SIGIR Conference on Research and Development in Information Retrieval*, volume 27, pages 564–565. ACM SIGIR, ACM Press, July 2004.

[23] Philip J. Cowans and Martin Szummer. A graphical model for simultaneous partitioning and labeling. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[24] R. T. Cox. *The Algebra Of Probable Inference*. John Hopkins University Press, 1961.

[25] A. P. Dawid. Conditional independence for statistical operations. *Annals Of Statistics*, 8:598–617, 1980.

[26] S. Deerwester, S. Dumais, T. Laundauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal Of The American Society Of Information Science*, 41(6):391–407, 1990.

[27] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *Annals Of Statistics*, 1(2), 1973.

[28] S. Goldwater, Thomas L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power law generators. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[29] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Scioences*, 101:5228–5235, 2004.

[30] Antonio Gulli and Alessio Signorini. The indexable web is more than 11.5 billion pages. In *Proceedings of the 14th International World Wide Web conference*, 2005.

[31] Djoerd Hiemstra and Wessel Kraaij. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *Text REtrieval Conference*, pages 174–185, 1998.

[32] T. Hofman. Probabilistic latent semantic indexing. In *Proceedings of the twenty-second annual SIGIR conference*, 1999.

[33] R. Nigel Horspool and Gordon V. Cormack. Constructing word-based text compression algorithms. In *Proceedings of the Data Compression Conference*, 1992.

[34] Edwin. T. Jaynes. *Probability Theory, The Logic Of Science*. Cambridge University Press, 2003.

[35] H. Jeffreys. *Theory Of Probability*. Oxford University Press, 1939. 3rd edition reprinted 1985.

[36] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[37] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, 1980.

[38] Finn V. Jensen. *Bayesian Networks And Decision Graphs*. Springer Verlag, 2001.

[39] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.

[40] Slava M. Katz. Estimation of probabiliies from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[41] Mark D. Kerninghan, Kenneth W. Church, and William A. Gale. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th Conference on Computational Linguistics.*, volume 2, pages 205–210, 1990.

[42] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal Of The ACM*, 46, 1999.

[43] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*, 2004.

[44] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram langauge modeling. In *Proceedings of the IEEE International Conference on Acoutics, Speech and Signal Processing*, volume 1, pages 131–184, 1995.

[45] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algroithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[46] S Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[47] J. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In J. Lafferty W. Bruce Croft, editor, *Langauge Modeling for Information Retrieval*, pages 1–11. Kluwer Academic Publishers, 2003.

[48] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conf. on Machine Learning*, 2001.

[49] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *2001 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, 2001.

[50] P. S. Laplace. *Philospohical Essays On probabilities*. Springer–Verlag, 1995. Originally published 1825.

[51] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[52] G. Lidstone. Note on the general case of the Bayes–Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, 1920.

[53] Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM)*, pages 375–382, 2002.

[54] Xiuwen Liu and DeLiang Wang. Perceptual organization based on temporal dynamics. In *Advances in Neurological Information Processing Systems*, volume 12, 2000.

[55] D. J. C. MacKay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1994.

[56] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[57] Chris Manning and Heinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[58] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the International Conference on Computer Vision*, pages 416–425, 2001.

[59] R. J. McEliece, D. J. C. MacKay, and J. F. Cheng. Turbo decoding as and instance of Pearl's 'belief propagation' algorithm. *IEEE Journal On Selected Areas In Communication*, 16(2):140–152.

[60] Alastair Moffat. A note on the PPM data compression algorithm. Technical Report 88/7, University of Melbourne, 1988.

[61] Alistair Moffat. Word-based text compression. *Software–Practice And Experience*, 19(2):185–198, 1989.

[62] Arthur Nádas. On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6):1414–1416, 1985.

[63] Herman Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[64] Paul Ogilive and Jamie Callan. Experiments using the Lemur toolkit. In *Proceedings of the Text REtrieval Conference*, 2002.

[65] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

[66] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *1998 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, 1998.

[67] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[68] Yuan Qi, Martin Szummer, and Thomas P. Minka. Diagram structure recognition by *b*ayesian conditional random fields. In *Proceedings of the International Conference on Computer Vision*, 2005.

[69] K. Srihari R, C. Ng, C. Baltus, and J. Kud. Use of language models in on-line recognition of handwritten sentences. In *Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition*, pages 284–294, 1993.

[70] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[71] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–87, 1986.

[72] S. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal Of The American Society For Information Science*, 27:129–146, 1976.

[73] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[74] S. E. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[75] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[76] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[77] J. Rocchio. Relevance feedback information retrieval. In Gerald Salton, editor, *The SMART retrieval system — Experiments in automatic document processing*, pages 313–323. Prentice-Hall, 1971.

[78] Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. In *Computer Speech And Language*, volume 10, pages 187–228, 1996.

[79] Ronald Rosenfeld. Two decades of statistical language modelling: Where do we go from here? *Proceedings Of The IEEE*, 88(8), 2000.

[80] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian approach to filtering junk E-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, 1998.

[81] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[82] E. Saund. Finding perceptually closed paths in sketches and drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):475–491, 2003.

[83] E. Saund, D. Fleet, D. Larner, and J. Mahoney. Perceptually-supported image editing of text and graphics. In *Proc. UIST 03*, pages 183–192, 2003.

[84] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, (4), 1994.

[85] F. Sha and F. Pereira. Parsing with conditional random fields. Technical Report MS-CIS-02-35, University of Pennsylvania, 2003.

[86] Claude E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, pages 50–64, 1951.

[87] P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence*, volume 4, pages 169–198, 1990.

[88] Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *International Conference on Machine Learning*, 2004.

[89] R. H. Swendsen and J. S. Wang. Nonuniversal critical dynamics in mc simulations. *Physical Review Letters*, 58(2):86–88, 1987.

[90] Martin Szummer and Philip J. Cowans. Incorporating context and user feedback in pen-based interfaces. In R. Davis and J. Landay et al., editors, *AAAI Fall symposium, Making Pen-Based Interaction Intelligent and Natural*, FS-04-06, pages 159–166. AAAI Press, 2004.

[91] Yee Whye Teh. A Bayesian interpretation of interpolated Kneser–Ney (nips workshop presentation). 2005.

[92] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. Technical Report 653, Department Of Statistics, UC Berkeley, 2003.

[93] Keith Vertanen. Efficient computer interfaces using continuous gestures, language models, and speech. Master's thesis, University of Cambridge, 2004.

[94] Hanna M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania, 2004.

[95] David Ward. *Adaptive Computer Interfaces*. PhD thesis, University of Cambridge, 2001.

[96] David J. Ward, Alan F. Blackwell, and David J. C. MacKay. Dasher — a data entry interface using continuous gestures and language models. In *UIST 2000: The 13th Annual ACM Symposium on User Interface Software and Technology*, 2000.

[97] David J. Ward and Davic J. C. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418:838, 2002.

[98] I. H. Witten, R. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

[99] Ian H. Witten and Timothy C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.

Web References

[100] A001861 from the on-line encyclopedia of integer sequences. `http://www.research.att.com/projects/OEIS?Anum=A001861`.

[101] Google book search. `http://books.google.com/`.

[102] Project Guttenberg. `http://www.gutenberg.org/`.

[103] University of Glasgow information retrieval resources. `http://ir.dcs.gla.ac.uk/resources/`.

[104] The Canterbury corpus. `http://corpus.canterbury.ac.nz`.

[105] The Dasher project. `http://www.dasher.org.uk/`.

[106] GNU Emacs website. `http://www.gnu.org/software/emacs/emacs.html`.

[107] Google. `http://www.google.com/`.

[108] Ispell spell checker. `http://www.gnu.org/software/ispell/ispell.html`.

[109] Mozilla Firefox website. `http://www.mozilla.org/products/firefox/`.

[110] MSN Search. `http://search.msn.com/`.

[111] *talks.cam* seminar database. `http://talks.cam.ac.uk/`.

[112] Tegic Communications, developers of the T9 text entry system. `http://www.tegic.com`.

[113] Text REtrieval Conferenec. `http://trec.nist.gov/`.

[114] Yahoo! `http://www.yahoo.com/`.