Polonius: What do you read, my lord?

Hamlet: Words, words, words.

Polonius: What is the matter, my lord?

Hamlet: Between who?

Polonius: I mean, the matter that you read, my lord.

Hamlet: Slanders, sir: for the satirical rogue says here that old men have grey beards....

Polonius: [*Aside*] Though this be madness, yet there is method in't.

–*Hamlet*, Act II, Scene ii.

What is the matter?

Text categorization (broadly construed): identification of "similar" documents.

Similarity criteria include:

- **topic** (subject matter)
- **source** (authorship or genre identification)
- **relevance** to a query (ad hoc information retrieval)
- **sentiment polarity**, or author's overall opinion (data mining)

Method to the madness

Syntax and semantics are ultimately necessary, but "**bag-of-words**"-based feature vectors are quite effective.

Can we do even better within a knowledge-lean framework?

Act I: Iterative Residual Re-scaling: a generalization of Latent Semantic Indexing (LSI) that creates <u>improved representations</u> for topic-based categorization

Act II: Sentiment analysis via minimum cuts: optimal incorporation of <u>pair-wise relationships</u> in a more semantically-oriented task

Joint work with Rie Kubota Ando (I) and Bo Pang (II).

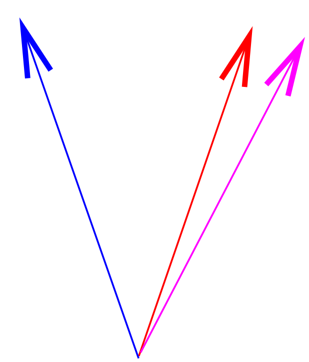
Words, words, words

Documents:

make hidden Markov model probabilities normalize	car emissions hood make model trunk	car engine hood tires truck trunk

Term−document matrix D:

$$
\begin{array}{cl}
\begin{bmatrix}
0 & 1 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 1 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 1 & 1
\end{bmatrix}
&
\begin{array}{l}
\text{car} \\
\text{emissions} \\
\text{engine} \\
\text{hidden} \\
\text{hood} \\
\text{make} \\
\text{Markov} \\
\text{model} \\
\text{normalize} \\
\text{probabilities} \\
\text{tires} \\
\text{truck} \\
\text{trunk}
\end{array}
\end{array}
$$



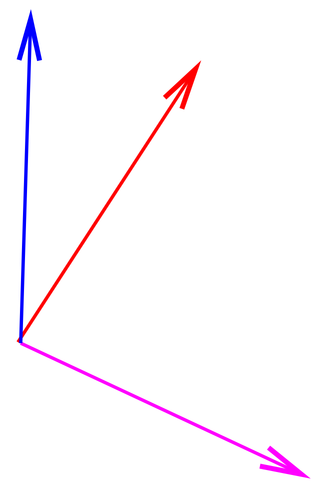
Problem: Synonymy

Documents:

make	car	auto
hidden	emissions	engine
Markov	hood	bonnet
model	make	tyres
probabilities	model	lorry
normalize	trunk	boot

Term–document matrix D:

$$
\begin{array}{l}
\left[\begin{array}{ccc}
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right]
\end{array}
\begin{array}{l}
\text{auto} \\
\text{bonnet} \\
\text{boot} \\
\text{car} \\
\text{emissions} \\
\text{engine} \\
\text{hidden} \\
\text{hood} \\
\text{lorry} \\
\text{make} \\
\text{Markov} \\
\text{model} \\
\text{normalize} \\
\text{probabilities} \\
\text{tires} \\
\text{trunk} \\
\text{tyres}
\end{array}
$$



Approach: Subspace projection

Project the document vectors into a lower-dimensional subspace.

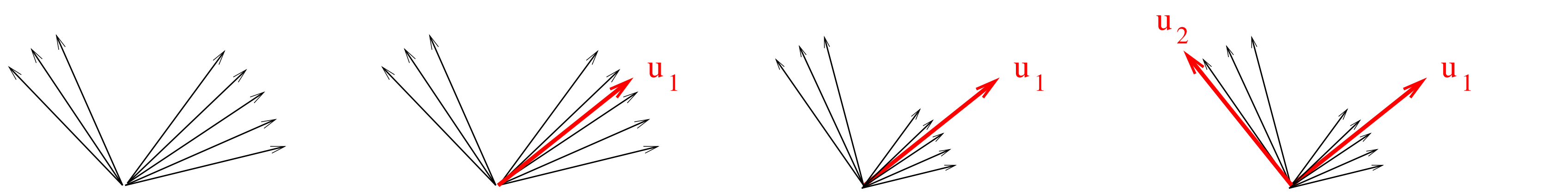
$\triangleright$ Synonyms no longer correspond to orthogonal vectors, so topic and directionality may be more tightly linked.

Most popular choice: Latent Semantic Indexing Deerwester et al. (1990) has $\sim$ 2200 citations:

- Pick some number k that is smaller than the rank of the term-document matrix D.
- Compute the first k *left singular vectors* $u_1, u_2, \ldots, u_k$ of D.
- .Set $\mathbf{D}'$ to the projection of $\mathbf{D}$ onto $\mathrm{span}(\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k})$.

Motivation: D' is the two-norm-optimal rank-k approximation to D (Eckart & Young, 1936).

A geometric view



Start with document vectors

Choose direction u maximizing projections

Compute ***residuals*** (subtract projections)

Repeat to get next u (orthogonal to previous u_i's)

That is, in each of k rounds, find

$$u = \arg\max_{v:|v|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(v, r_j)) \qquad \text{(“weighted average”)}$$

But, is the induced optimum rank-k approximation to the original term-document matrix *also* the optimal representation of the documents?

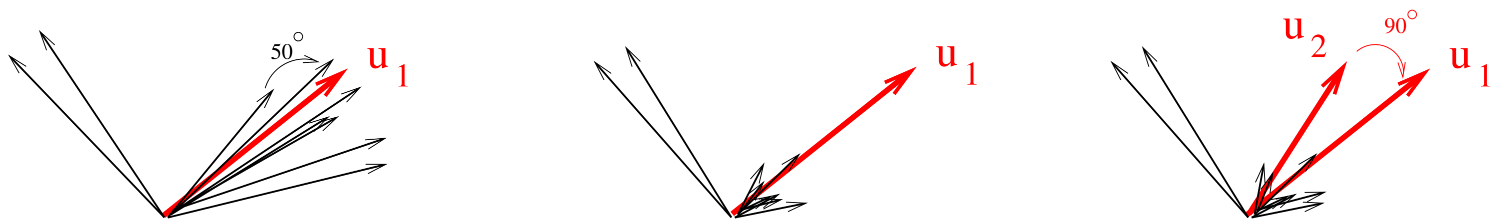
Results are mixed; e.g., Dumais et al. (1998).

Arrows of outrageous fortune

Recall: in each of k rounds, LSI finds

$$u = \arg\max_{v:|v|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(v, r_j))$$

Problem: Non-uniform distributions of topics among documents



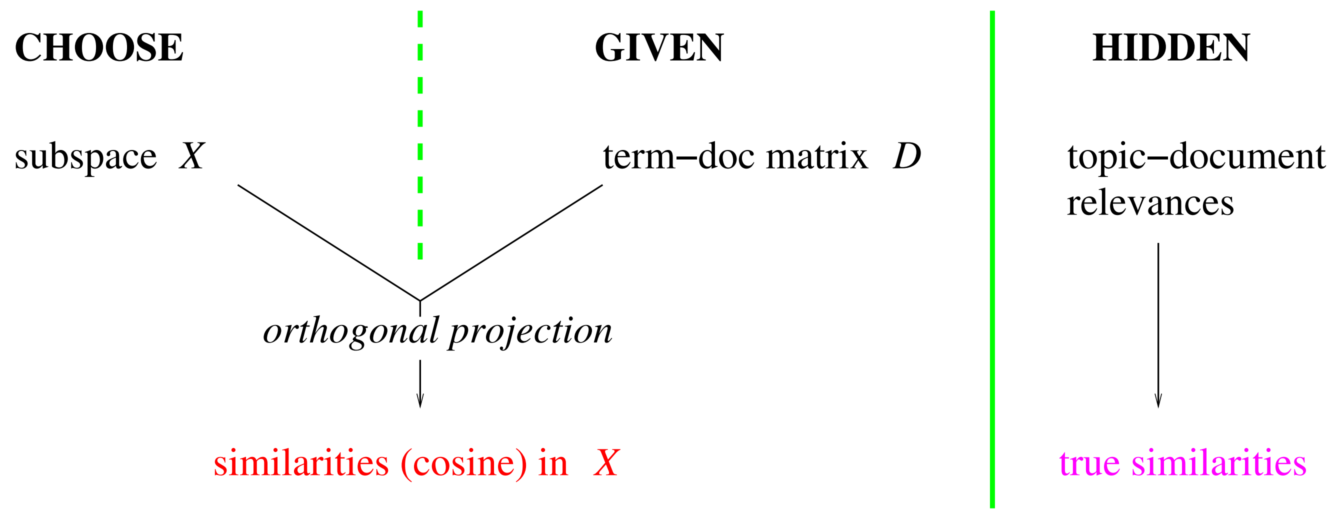
Choose direction u maximizing projections

Compute residuals

Repeat to get next u (orthogonal to previous u_i's)

dominant topics bias the choice

Gloss of main analytic result



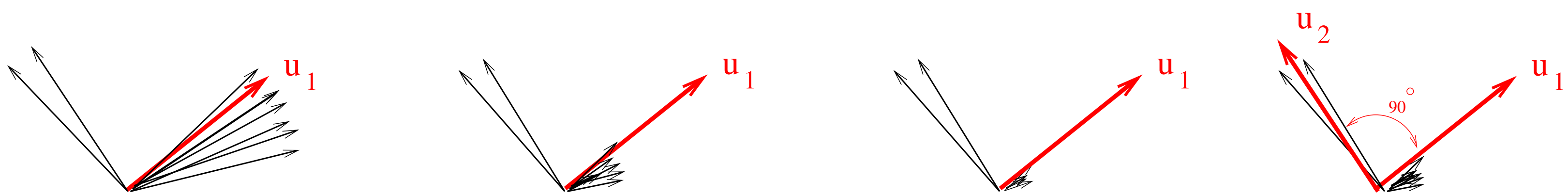
Under mild conditions, the distance between X^{LSI} and $X^{optimal}$ is bounded by a function of the topic-document distribution's non-uniformity and other reasonable quantities, such as D's "distortion".

Cf. analyses based on generative models (Story, 1996; Ding, 1999; Papadimitriou et al., 1997, Azar et al., 2001) and empirical comparison of X^{LSI} with an optimal subspace (Isbell and Viola, 1998).

By indirections find directions out

Recall: $u = \arg\max_{v:|v|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(v, r_j))$

We can compensate for non-uniformity by re-scaling the residuals by the sth power of their length at each iteration: $r_j \longrightarrow |r_j|^s \cdot r_j$ (Ando, 2000).



Choose direction u maximizing projections | Compute residuals | Rescale residuals (relative diffs rise) | Repeat to get next u (orthogonal to previous u_i's)

The **Iterative Residual Re-scaling** algorithm (IRR) estimates the (unknown) non-uniformity to *automatically* set the scaling factor s

Later work (Cristianini, Shawe-Taylor, & Lodhi 2002): supervised re-scaling

Experimental framework

We used TREC documents. Topic labels served as validation.

Stop-words removed; no term weighting.

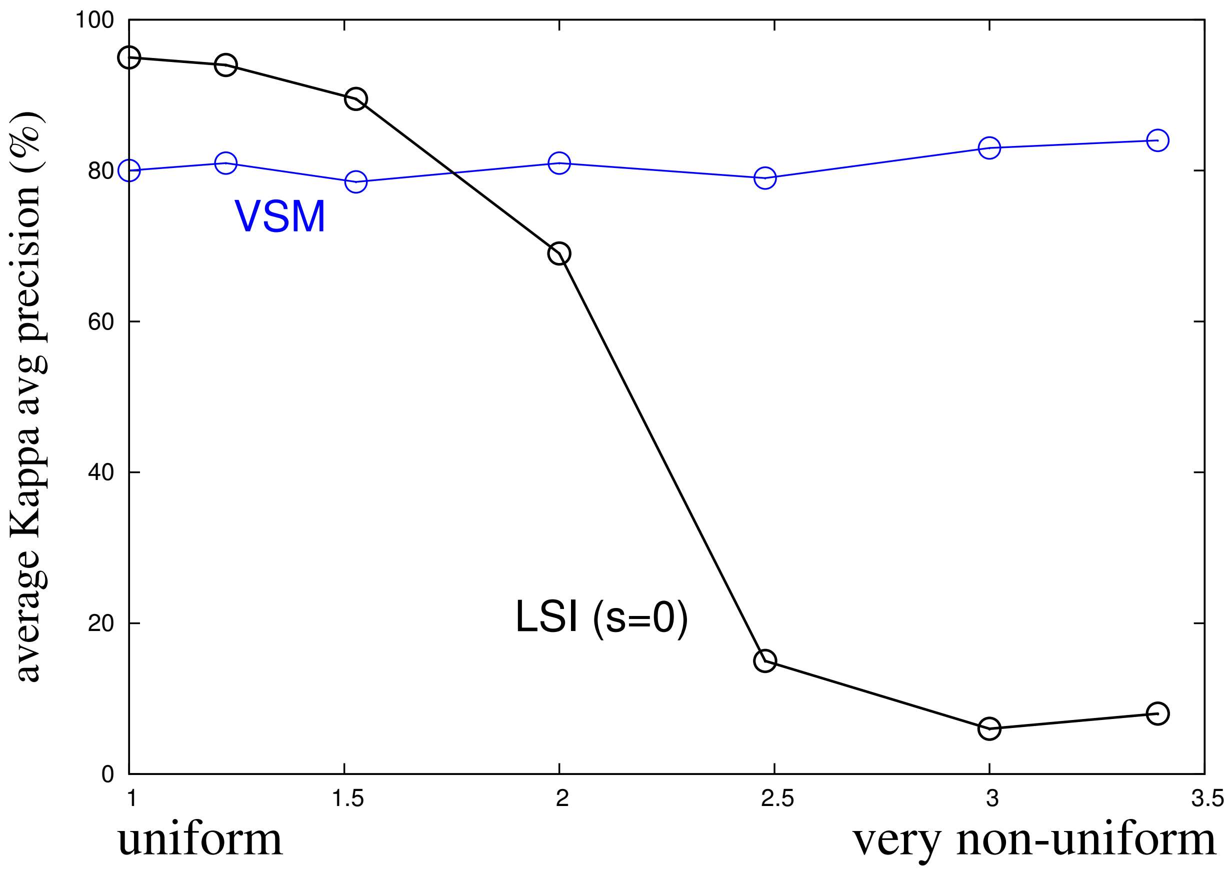
Scoring function (*not* algorithm) requires single-topic documents.

Controlled distributions: we manually altered topic dominances to study their effects on LSI and IRR's performance.

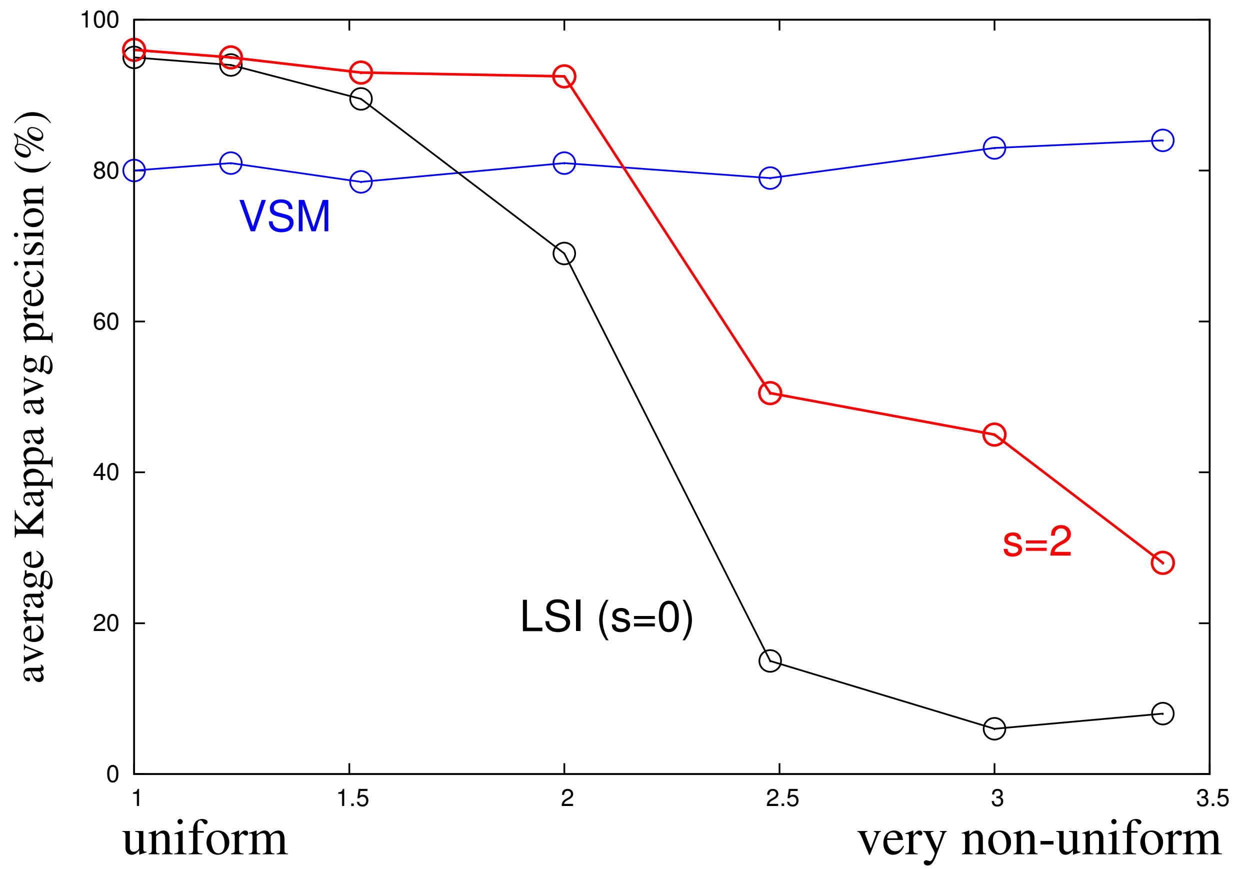
- For a set of k topics, for a sequence of increasingly non-uniform distributions, ten 50-document sets for each such distribution were created. (The subspace dimensionality was fixed at k.)

Uncontrolled distributions: we simulated retrieval results.

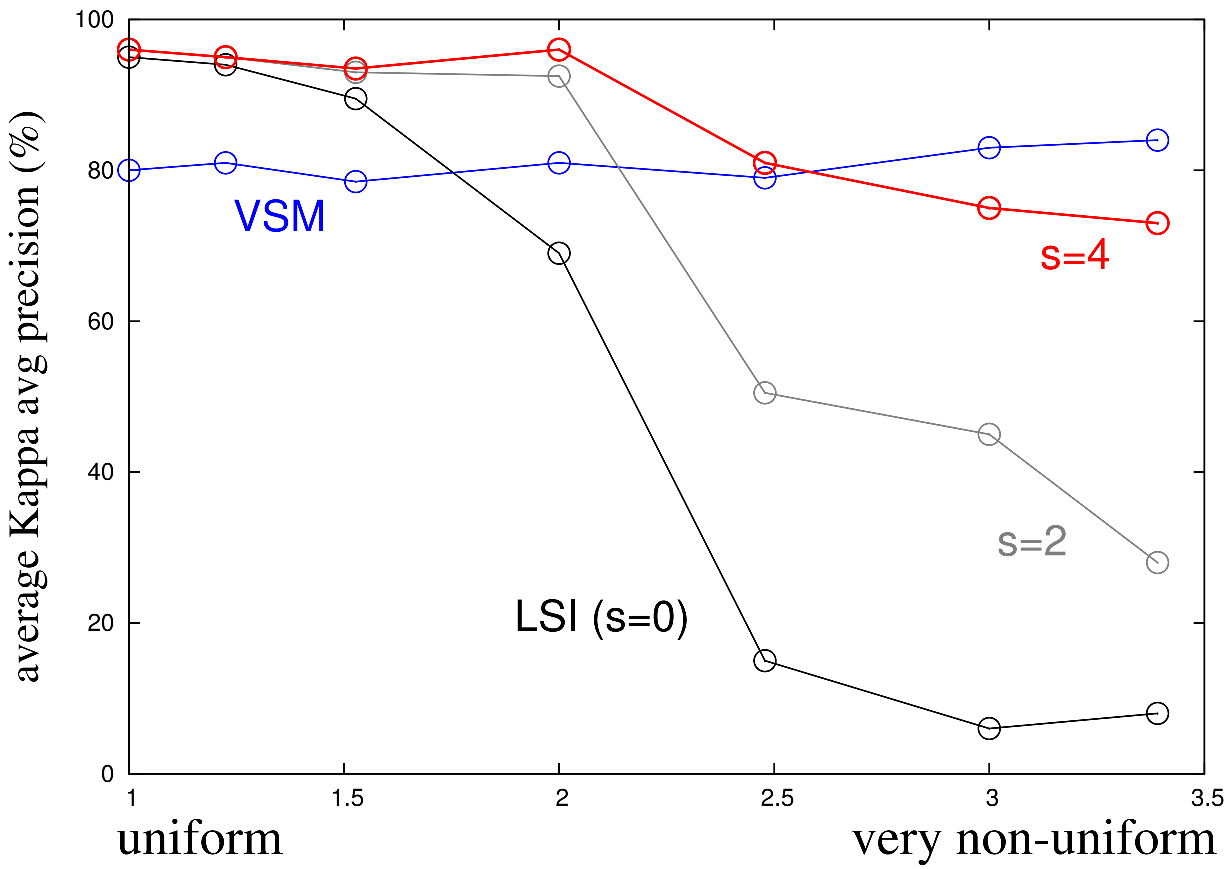
- For each keyword in a randomly-chosen set of 15, all documents containing that keyword were selected to create a document set.



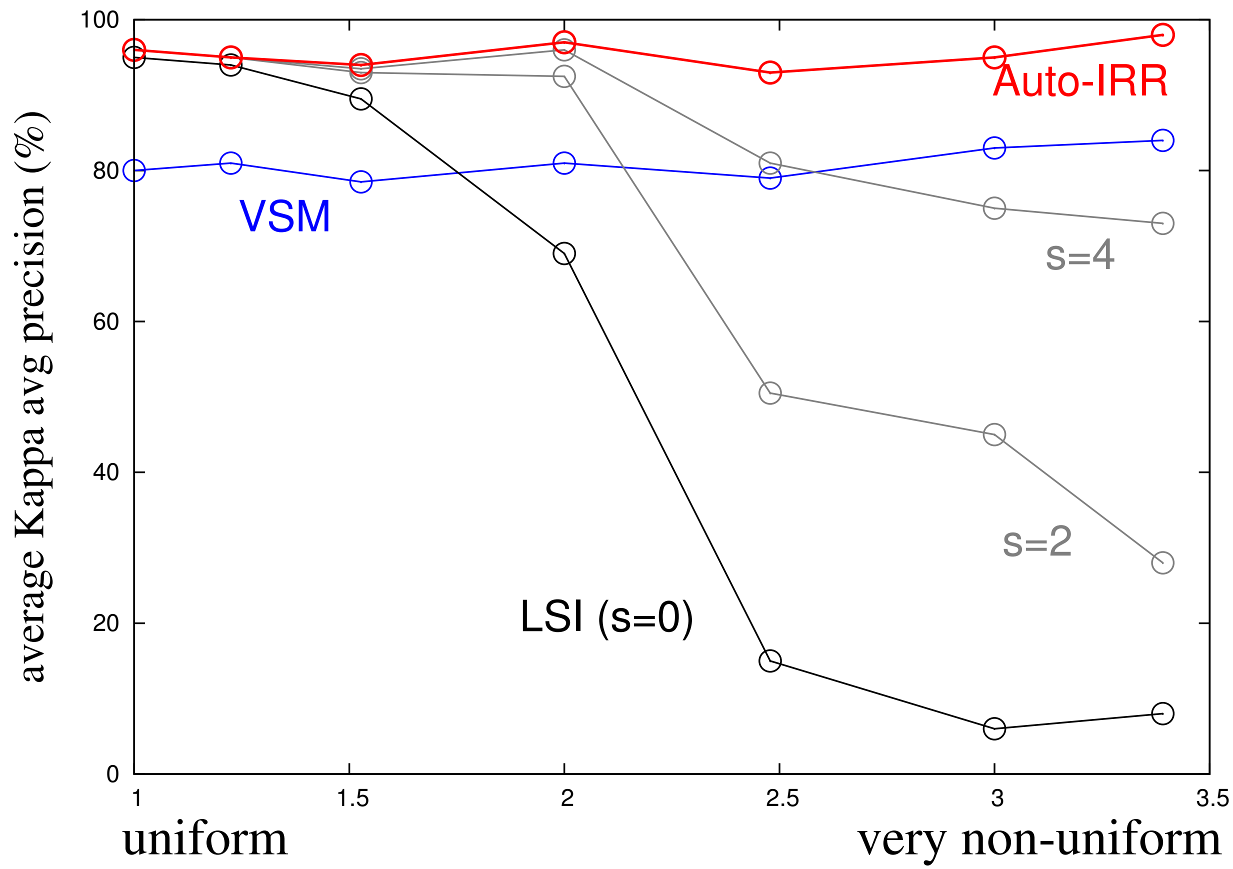
Each point: average over 10 different datasets of the given non-uniformity.



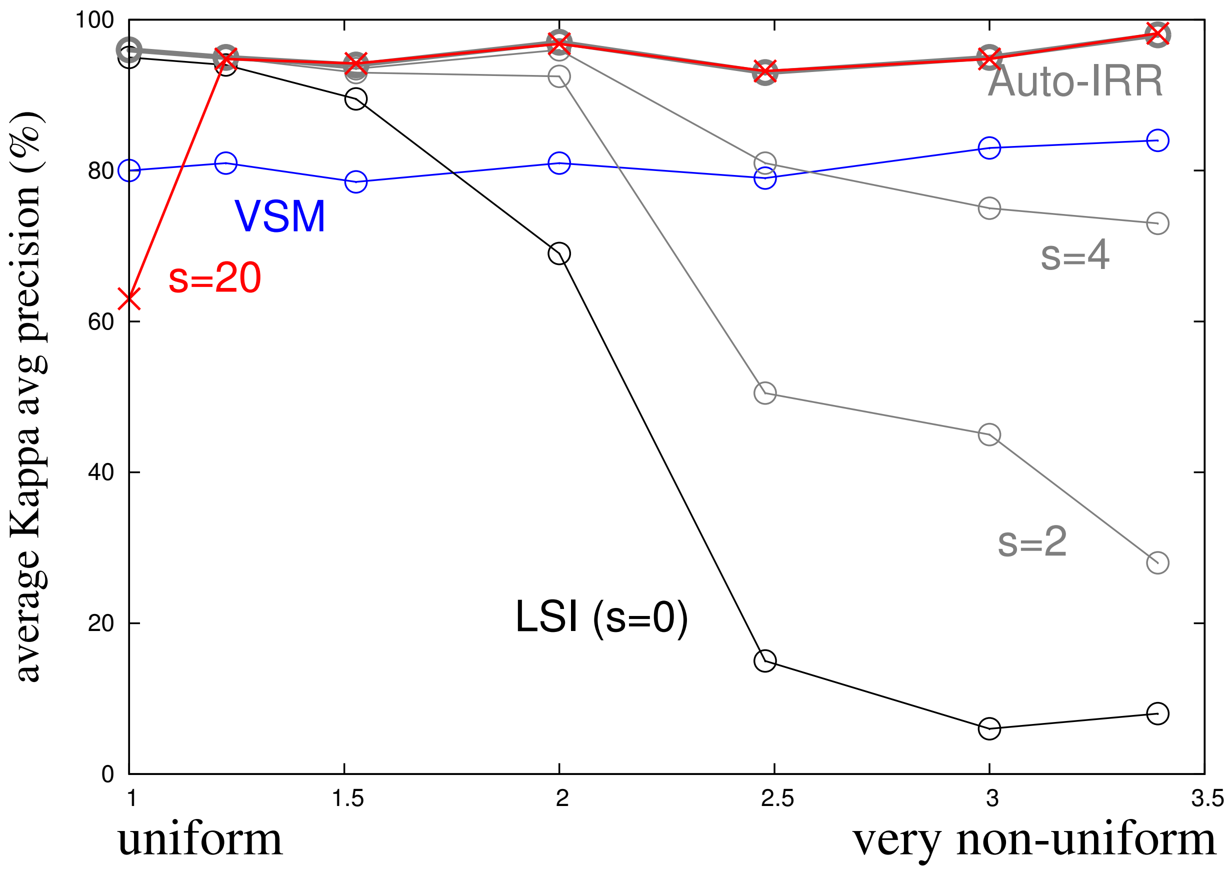
Each point: average over 10 different datasets of the given non-uniformity.



Each point: average over 10 different datasets of the given non-uniformity.

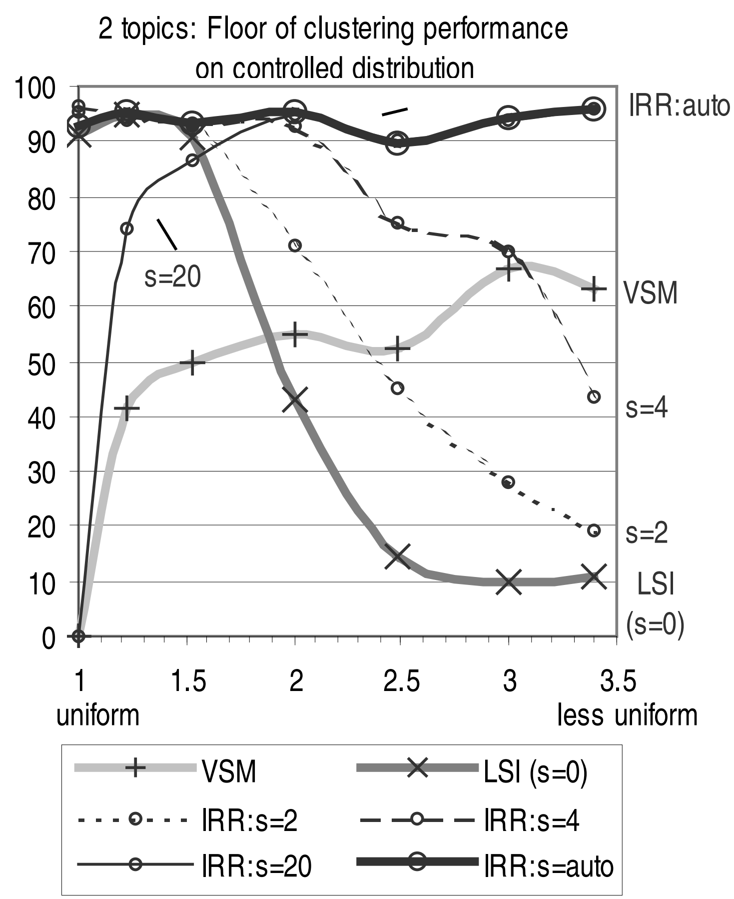


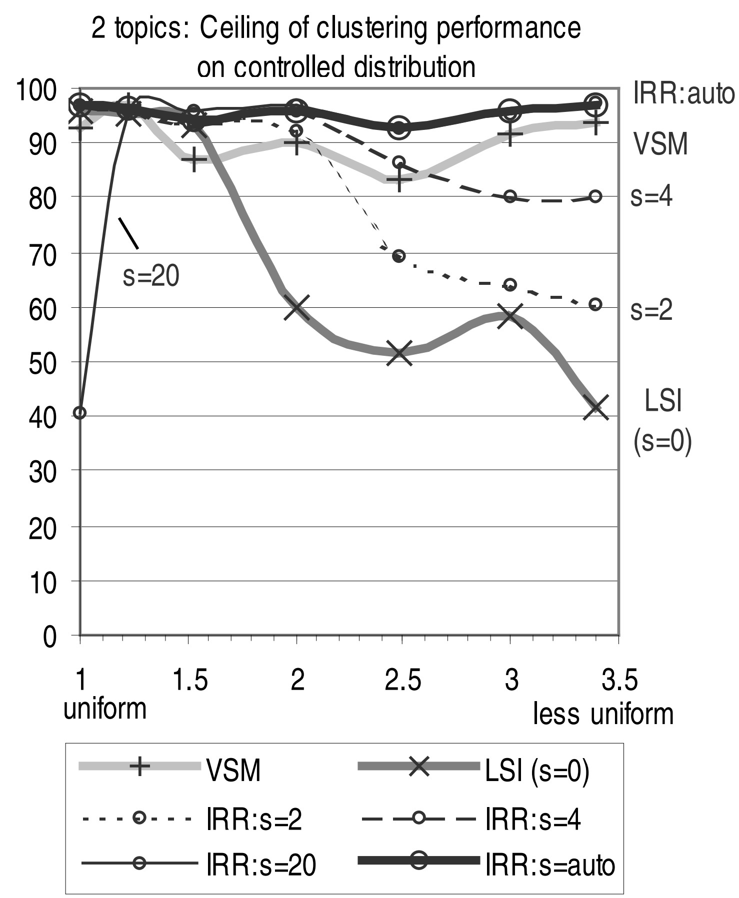
Each point: average over 10 different datasets of the given non-uniformity.



Each point: average over 10 different datasets of the given non-uniformity.

Clustering results





Act II: Nothing either good or bad, but thinking makes it so

We've just explored improving text categorization based on *topic*.

An interesting alternative: sentiment polarity — an author's overall opinion towards his/her subject matter ("thumbs up" or "thumbs down").

Applications include:

- providing summaries of reviews, feedback, blog posts, and surveys
- organizing opinion-oriented text for IR or question-answering systems

General sentiment analysis is the computational treatment of subjective or opinionated text (Wiebe 1994; Das & Chen 2001; Pang, Lee & Vaithyanathan 2002; Turney 2002; Dave, Lawrence & Pennock 2003; Yu & Hatzivassiloglou 2003); applications include generation (Inkpen, Feiguina & Hirst 2005) and medical informatics (Niu, Zhu, Li, & Hirst 2005).

More matter, with less art

State-of-the-art methods using bag-of-words-based feature vectors have proven less effective for sentiment classification than for topic-based classification (Pang, Lee, and Vaithyanathan, 2002).

- This laptop is <u>a great deal</u>.
- <u>A great deal</u> of media attention surrounded the release of the new laptop.
- If you think this laptop is <u>a great deal</u>, I've got a nice bridge you might be interested in.
- This film should be <u>brilliant</u>. It sounds like a <u>great</u> plot, the actors are <u>first grade</u>, and the supporting cast is <u>good</u> as well, and Stallone is attempting to deliver a <u>good</u> performance. However, it can't hold up.

From: $\langle$somebody@cs.somewhere.edu$\rangle$

Date: Thu, 8 Sep 2005 08:47:03 -0400

Subject: FW: [Colloquium-l] Reminder - Lillian Lee speaks TODAY

[*snip*]

The topic sounds very interesting. Frankly, I'm skeptical! And I bet your analysis won't figure that out just from this email, either...

Brevity is the soul of wit

We propose employing sentiment summarization on reviews:

1. Identify and retain only the *subjective* sentences.

2. Classify the induced subjectivity extract instead of the full review.

This yields another classification problem (on sentences as subjective or objective), but some advantages are:

- objective portions, such as background material or plot descriptions, may contain misleading text ("A great deal of media attention ...")

- users can use the extracts as summaries

Incorporating sentence relationships

Example: *Two sentences that are close together tend to have the same subjectivity status, unless separated by paragraph boundaries.*

Given instances $x_1, \ldots, x_n$, labels C_1 and C_2, and (non-negative) ...

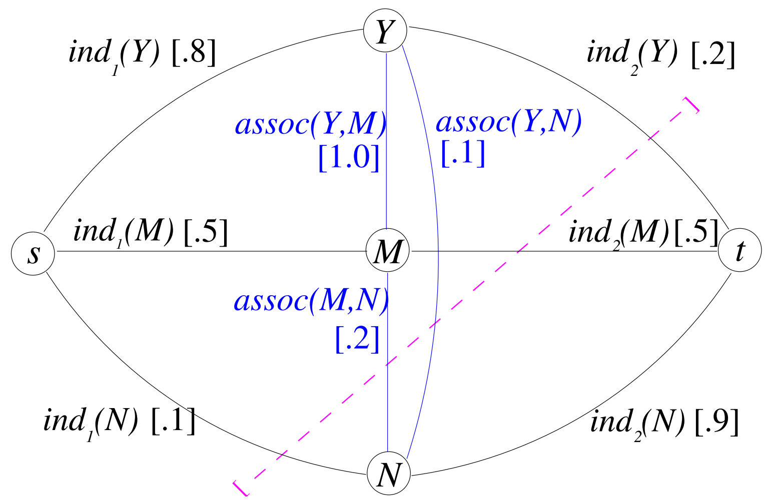
- individual class-preference scores for each x_i for each C_k, and
- association scores for each (x_i, x_j) pair,

we desire a labeling that minimizes

$$\sum_{x \in C_1} ind_2(x) + \sum_{x \in C_2} ind_1(x) + \sum_{\substack{x_i \in C_1, \\ x_k \in C_2}} assoc(x_i, x_k),$$

or, equivalently, maximizes each x_i's individual net happiness with its assigned class and with its class-mates.

Graph formulation and minimum cuts



Each labeling corresponds to a partition, or cut, whose cost is the sum of weights of edges with endpoints in different partitions (for symmetric assoc.).

Using network-flow techniques, computing the minimum cut...

- takes polynomial time, worst case, and little time in practice (Ahuja, Magnanti, & Orlin, 1993)
- special case: finding the maximum a posteriori labeling in a Markov random field (Besag 1986; Greig, Porteous, & Seheult, 1989)

Related work

Previous applications of the min-cut paradigm: vision (Greig, Porteous, &Seheult, 1989; Boykov, Veksler, & Zabih, 1999; Boykov & Huttenlocher, 1999; Kolmogorov & Zabih, 2002; Raj & Zabih, 2005); computational biology (Kleinberg 1999; Xu, Xu, & Gabow, 2000; Aspnes et al, 2001); Web mining (Flake, Lawrence, & Giles, 2000); learning with unlabeled data (Blum & Chawla 2001)

Later applications of minimum-cut-based methods in NLP: sentiment analysis (Pang & Lee 2005; Agarwal & Bhattacharyya 2005; Thomas, Pang, & Lee 2006), generation (Barzilay & Lapata 2005)

Examples of other methods incorporating relationship information: Graph partitioning (Shi & Malik, 1997; Bansal, Blum, & Chawla, 2002; Joachims, 2003; Malioutov & Barziley 2006) probabilistic relational models and related “collective classification” formalisms (Friedman et al., 1999; Lafferty, McCallum, & Pereira 2001; Neville and Jensen, 2003; Taskar et al. 2004)

Evaluation framework

Data:

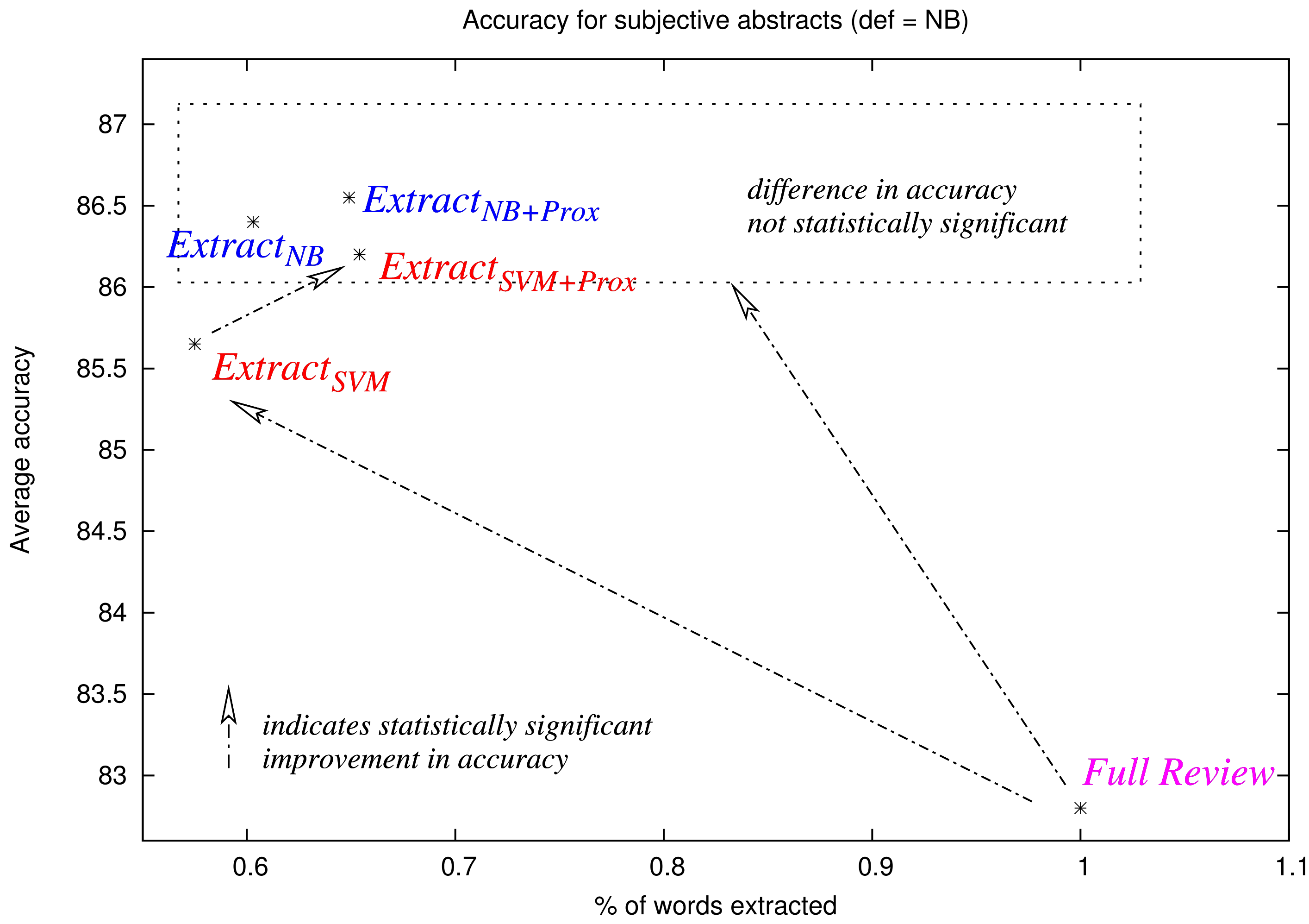
- 1000 positive and 1000 negative reviews from the IMDb, pre-2002
  - ▷ True labels extracted from rating info (e.g., "$\star$ out of $\star\star\star\star$")
- 5000 subjective sentences: "snippets" from Rotten Tomatoes, post-2001
- 5000 objective sentences: IMDb plot-summary sentences, post-2001

See http://www.cs.cornell.edu/people/pabo/movie-review-data .

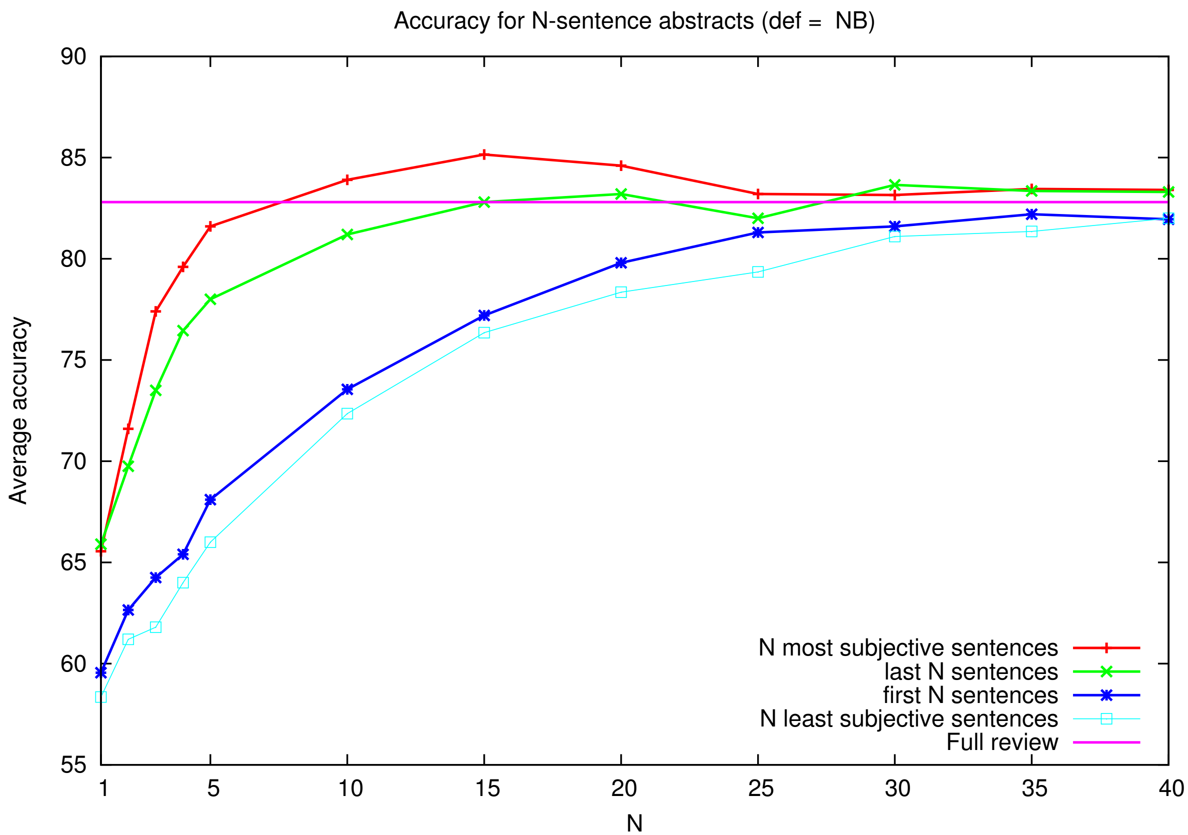
All results: ten-fold cross-validation average performance; paired t-test for significance testing.

(The extract classifiers and individual-score sentence classifiers (Naive Bayes and SVMs for both) were trained; we left association-score parameter selection for future work.)

(One set of) summarization results



(One set of) “top n”-sentence results



Shortest review: “This film is extraordinarily horrendous and I’m not going to waste any more words on it”.

The undiscovered country

We discussed:

- Better choice of feature vectors for document representation via IRR
  - ▷ Bounds analogous to those for LSI on IRR?
  - ▷ Alternative ways to compensate for non-uniformity?
- Incorporation of pairwise coherence constraints into subjectivity summarization using a minimum-cut paradigm
  - ▷ Other constraints, either knowledge-lean or knowledge-based?
  - ▷ Transductive learning for selecting association-constraint parameters?
  - ▷ Other applications?
    - ∗ Example: Thomas, Pang & Lee (2006): classifying political speeches as supporting/opposing legislation, using indications of agreement between speakers to create links