

# Beyond Document Retrieval

Maarten de Rijke

Language & Inference Technology group

ILLC, U. of Amsterdam

<http://lit.science.uva.nl>



Veb

Imeges

Gruops

Durectury

deffoolt seerch behefeur

Google-a Seerch

Ee'm Feeleeng Looky

- [Edfunced Seerch](#)
- [Preferences](#)
- [Lunqooege-a Tuuls](#)

[Eil Ebuoot Google-a - Google in English](#)

©2003 Google - Seercheeng 3,083,324,652 veb pages

# Motivation

# Motivation

- ▶ Modern search engines such as Google provide extensive **document retrieval**

# Motivation

- ▶ Modern search engines such as Google provide extensive **document retrieval**
- ▶ What does this mean?

# Motivation

- ▶ Modern search engines such as Google provide extensive **document retrieval**
- ▶ What does this mean?
- ▶ Given a collection of documents, and a number of keywords, the retrieval engine returns a ranked list of relevant documents

# Motivation

- ▶ Modern search engines such as Google provide extensive **document retrieval**
- ▶ What does this mean?
- ▶ Given a collection of documents, and a number of keywords, the retrieval engine returns a ranked list of relevant documents
- ✓ Efficient access to '3,083,324,652 web pages'

# Wish List

# Wish List

- ▶ Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear

# Wish List

## ► Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear
- ✗ . . . but I have a collection of XML documents, and I want to exploit the document structure

# Wish List

## ► Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear
- ✗ . . . but I have a collection of XML documents, and I want to exploit the document structure
- ✗ . . . but I have multi-lingual docs and I want relevance across language boundaries

# Wish List

## ▶ Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear
- ✗ . . . but I have a collection of XML documents, and I want to exploit the document structure
- ✗ . . . but I have multi-lingual docs and I want relevance across language boundaries

## ▶ Higher level

- ✗ . . . but I want the **really** relevant information

# Wish List

## ▶ Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear
- ✗ . . . but I have a collection of XML documents, and I want to exploit the document structure
- ✗ . . . but I have multi-lingual docs and I want relevance across language boundaries

## ▶ Higher level

- ✗ . . . but I want the **really** relevant information
- ✗ . . . but I have a question, and I want an **answer**, not a doc

# Wish List

## ▶ Technical

- ✗ . . . but I have a domain-specific collection to which I want to bring a lot of additional knowledge to bear
- ✗ . . . but I have a collection of XML documents, and I want to exploit the document structure
- ✗ . . . but I have multi-lingual docs and I want relevance across language boundaries

## ▶ Higher level

- ✗ . . . but I want the **really** relevant information
- ✗ . . . but I have a question, and I want an **answer**, not a doc
- ✗ . . . but what do people **think about** this information?

# Agenda for Today

# Agenda for Today

- ▶ A look at two ongoing (far from complete) research activities aimed at intelligent information access

# Agenda for Today

- ▶ A look at two ongoing (far from complete) research activities aimed at intelligent information access
  - **novelty**: pinpointing relevant information, including only what you have not seen before

# Agenda for Today

- ▶ A look at two ongoing (far from complete) research activities aimed at intelligent information access
  - **novelty**: pinpointing relevant information, including only what you have not seen before
  - **opinion extraction**: automatically extracting the subjective meaning of texts

# Agenda for Today

- ▶ A look at two ongoing (far from complete) research activities aimed at intelligent information access
  - **novelty**: pinpointing relevant information, including only what you have not seen before
  - **opinion extraction**: automatically extracting the subjective meaning of texts
- ▶ Intelligent information access requires a mixture of information retrieval, natural language processing, and artificial intelligence

# Agenda for Today

- ▶ A look at two ongoing (far from complete) research activities aimed at intelligent information access
  - **novelty**: pinpointing relevant information, including only what you have not seen before
  - **opinion extraction**: automatically extracting the subjective meaning of texts
- ▶ Intelligent information access requires a mixture of information retrieval, natural language processing, and artificial intelligence
- ▶ To begin, a slide or two on evaluation and TREC

# TREC

# TREC

- ▶ Assessing the quality of a information access systems requires objective evaluation

# TREC

- ▶ Assessing the quality of a information access systems requires objective evaluation
  - e.g., for a query and set of relevant docs check whether the system identified the relevant sentences without duplication

# TREC

- ▶ Assessing the quality of a information access systems requires objective evaluation
  - e.g., for a query and set of relevant docs check whether the system identified the relevant sentences without duplication
  - e.g., for a set of questions one has to check whether the system returns the correct answer. . . this is a laborious process

# TREC

- ▶ Assessing the quality of a information access systems requires objective evaluation
  - e.g., for a query and set of relevant docs check whether the system identified the relevant sentences without duplication
  - e.g., for a set of questions one has to check whether the system returns the correct answer. . . this is a laborious process
- ▶ TREC = Text REtrieval Conference
  - <http://trec.nist.gov>
  - TREC-1 held in 1991, . . . , TREC-11 in 2002
  - run by NIST, organized by IR research community
  - focus originally on “large” text collections (multi-Gb)

# TREC (2)

# TREC (2)

- ▶ TREC now includes many IR-related tracks
  - filtering, cross-language, multi-lingual, novelty, video retrieval, interactive, very large collections, Web, question answering, robust retrieval, . . .

# TREC (2)

- ▶ TREC now includes many IR-related tracks
  - filtering, cross-language, multi-lingual, novelty, video retrieval, interactive, very large collections, Web, question answering, robust retrieval, . . .
- ▶ What do you get from TREC?
  - document collections, query/question sets
  - scoring and evaluation methods
  - task definition
  - human assessors to judge the results returned

# Novelty Track

# Novelty Track

- ▶ The novelty track is one of several ongoing activities aimed at pinpointing relevant information

# Novelty Track

- ▶ The novelty track is one of several ongoing activities aimed at pinpointing relevant information
- ▶ Other/related pinpointing activities
  - XML retrieval (evaluated at INEX)
  - question answering (evaluated at TREC)

# Novelty Track

# Novelty Track

- ▶ Return only new **and** relevant sentences (within context) rather than whole documents containing duplicate and extraneous information

# Novelty Track

- ▶ Return only new **and** relevant sentences (within context) rather than whole documents containing duplicate and extraneous information
  - application scenario: a smart **next** button that walks a user down the ranked list of relevant documents and that hits the next new and relevant sentence

# Novelty Track

- ▶ Return only new **and** relevant sentences (within context) rather than whole documents containing duplicate and extraneous information
  - application scenario: a smart **next** button that walks a user down the ranked list of relevant documents and that hits the next new and relevant sentence
- ▶ TREC: novelty track
  - 50 topics
  - return two lists of doc id/sentence number pairs for each topic, one with the **relevant** sentences, and the second (a subset of the first) containing only sentences with **new** information

# Novelty Track

# Novelty Track

- ▶ 2002 was the first year the novelty track was held at TREC

# Novelty Track

- ▶ 2002 was the first year the novelty track was held at TREC
- ▶ Some topic/document details
  - the 50 topics are taken from TRECs 6, 7, and 8 (more below)
  - newspaper collection (late 1980s, early 1990s): LATimes, FBIS, Financial Times

# Novelty Track

- ▶ 2002 was the first year the novelty track was held at TREC
- ▶ Some topic/document details
  - the 50 topics are taken from TRECs 6, 7, and 8 (more below)
  - newspaper collection (late 1980s, early 1990s): LATimes, FBIS, Financial Times
- ▶ Our main interest: the **novelty** part, but first use established IR strategies for the relevance part

# Topic 414

# Topic 414

<top>

<num> Number: 414

<title> Cuba, sugar, exports

<desc> Description:

How much sugar does Cuba export and which countries import it?

<desc2> Description:

How much sugar does Cuba export and which countries import it?

<narr> Narrative:

A relevant document will provide information regarding Cuba's sugar trade. Sugar production statistics are not relevant unless exports are mentioned explicitly.

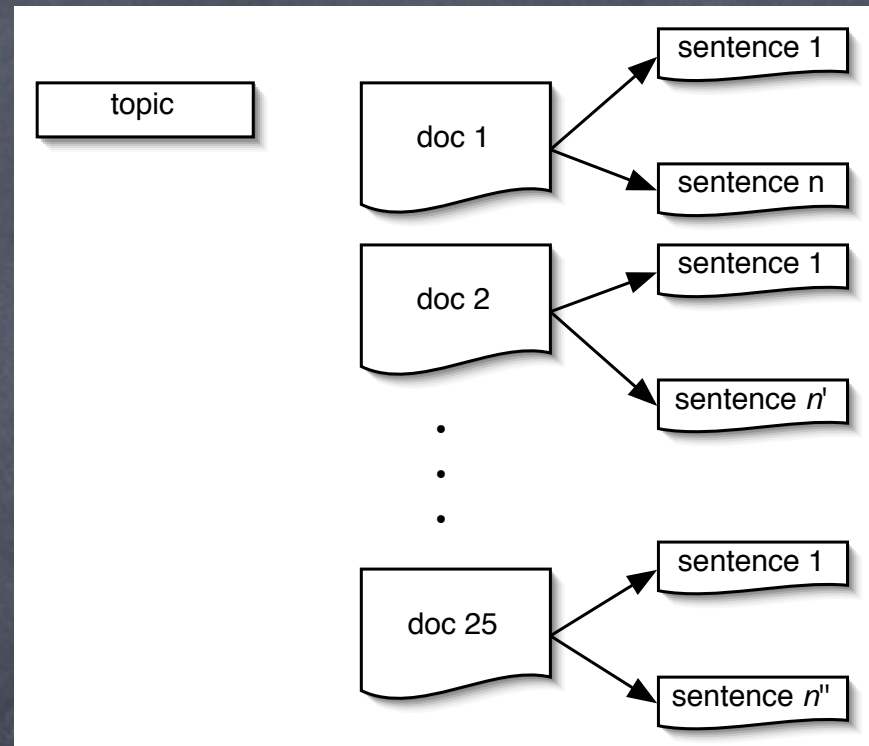
# The Relevance Part

# The Relevance Part

- ▶ For a given topic, view the sentences in the relevant documents for the topic as docs themselves

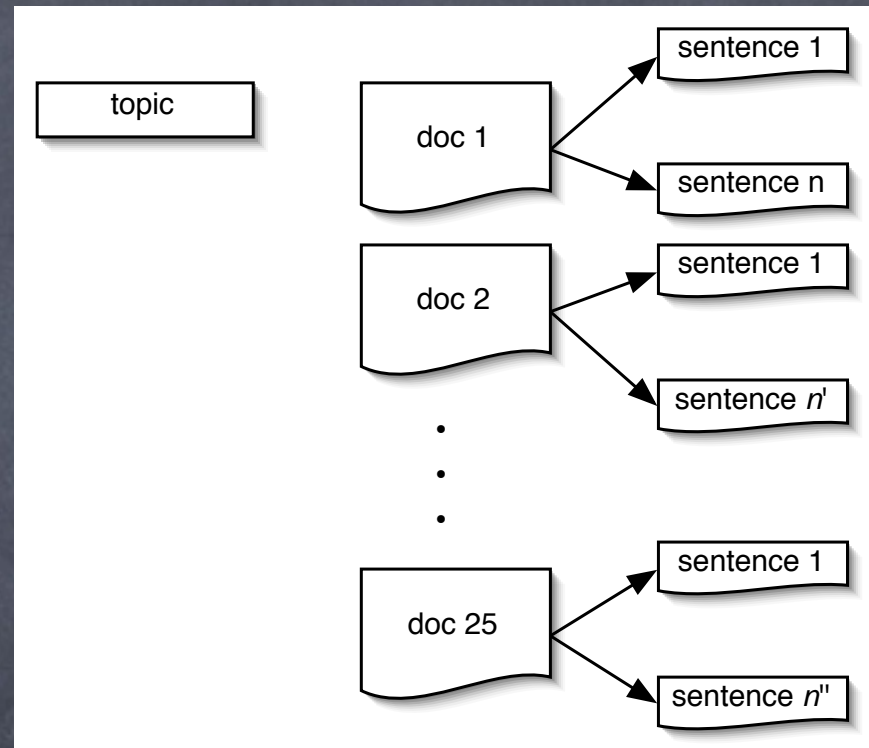
# The Relevance Part

- ▶ For a given topic, view the sentences in the relevant documents for the topic as docs themselves



# The Relevance Part

- ▶ For a given topic, view the sentences in the relevant documents for the topic as docs themselves
- ▶ Perform a traditional retrieval run: topic against this sentence-as-doc collection



# The Relevance Part (2)

## The Relevance Part (2)

- ▶ Re-order document id/sentence numbers by their occurrence in the original ranked list of docs

## The Relevance Part (2)

- ▶ Re-order document id/sentence numbers by their occurrence in the original ranked list of docs
- ▶ How do traditional doc retrieval ideas work here?
  - similarity, cut-off
  - topic and docs representation
  - . . .

## The Relevance Part (2)

- ▶ Re-order document id/sentence numbers by their occurrence in the original ranked list of docs
- ▶ How do traditional doc retrieval ideas work here?
  - similarity, cut-off
  - topic and docs representation
  - . . .
- ▶ How good does the relevance part need to be before something interesting can be done on the novelty part?

# Approaches

# Approaches

- ▶ Simple tf.idf based similarity
  - $\sim$  topic and docs (i.e., sentences) are viewed as vectors of terms, and similarity is the 'angle between the vectors'

# Approaches

- ▶ Simple tf.idf based similarity
  - $\sim$  topic and docs (i.e., sentences) are viewed as vectors of terms, and similarity is the 'angle between the vectors'
- ▶ Topic & document representation
  - word-based: use lower-cased words (with minor stopping)
  - stemmed: determine the root of a words, and use that
  - lemmatized: idem, but linguistically informed

# Approaches

- ▶ Simple tf.idf based similarity
  - $\sim$  topic and docs (i.e., sentences) are viewed as vectors of terms, and similarity is the 'angle between the vectors'
- ▶ Topic & document representation
  - word-based: use lower-cased words (with minor stopping)
  - stemmed: determine the root of a words, and use that
  - lemmatized: idem, but linguistically informed
- ▶ Which of the T, D, N fields to use?
  - experience from other retrieval tasks suggests using more, as this increases effectiveness

# Approaches (2)

# Approaches (2)

- ▶ Query expansion
  - move from strings to concepts, by adding synonyms (or more specific/general terms, or . . . ) to the original topic words

# Approaches (2)

- ▶ Query expansion
  - move from strings to concepts, by adding synonyms (or more specific/general terms, or . . . ) to the original topic words
  - synonyms etc from WordNet or through collocation

## Approaches (2)

- ▶ Query expansion
  - move from strings to concepts, by adding synonyms (or more specific/general terms, or . . . ) to the original topic words
  - synonyms etc from WordNet or through collocation
- ▶ Document expansion
  - topics are often general, documents specific
  - expand the **specific** terms in the docs, by adding more **general** terms if those more general terms occur in the topic

## Approaches (2)

- ▶ Query expansion
  - move from strings to concepts, by adding synonyms (or more specific/general terms, or . . . ) to the original topic words
  - synonyms etc from WordNet or through collocation
- ▶ Document expansion
  - topics are often general, documents specific
  - expand the **specific** terms in the docs, by adding more **general** terms if those more general terms occur in the topic
- ▶ Linguistic clues
  - first sentence, long sentences, cue phrases

# About the Test Set

# About the Test Set

- ▶ Human assessors created relevance and novelty lists
  - these were used as the 'golden standard'

# About the Test Set

- ▶ Human assessors created relevance and novelty lists
  - these were used as the 'golden standard'
- ▶ For each of 50 topics, the “minimum” assessor (the one who selected the fewest number of relevant sentences) was designated as the “official” assessor

# About the Test Set

- ▶ Human assessors created relevance and novelty lists
  - these were used as the 'golden standard'
- ▶ For each of 50 topics, the “minimum” assessor (the one who selected the fewest number of relevant sentences) was designated as the “official” assessor
- ▶ Using the minimum assessor, the median percentage of sentences marked relevant is about 2%
  - only 3 topics have more than 5% relevant sentences

# About the Test Set

- ▶ Human assessors created relevance and novelty lists
  - these were used as the 'golden standard'
- ▶ For each of 50 topics, the “minimum” assessor (the one who selected the fewest number of relevant sentences) was designated as the “official” assessor
- ▶ Using the minimum assessor, the median percentage of sentences marked relevant is about 2%
  - only 3 topics have more than 5% relevant sentences
- ▶ Median percentage of relevant sentences marked novel is 93%

# Scoring

# Scoring

- ▶ For both (relevance and novelty) lists **recall** and **precision** are used
  - $\text{recall} = |\textit{relevant matched}| / |\textit{relevant}|$
  - $\text{precision} = |\textit{relevant matched}| / |\textit{sentences submitted}|$
  - similarly for novelty, with novel 'instead' of 'relevant'

# Scoring

- ▶ For both (relevance and novelty) lists **recall** and **precision** are used
  - $\text{recall} = |\text{relevant matched}| / |\text{relevant}|$
  - $\text{precision} = |\text{relevant matched}| / |\text{sentences submitted}|$
  - similarly for novelty, with novel 'instead' of 'relevant'
- ▶ Single-value measure that averages well: **F measure**

$$\frac{2 * P * R}{P + R}$$

# Some Scores

# Some Scores

## Relevance scores (Avg F measure)

|                   |       |
|-------------------|-------|
| Human performance | 0.371 |
|-------------------|-------|

|        |       |
|--------|-------|
| Random | 0.040 |
|--------|-------|

|                   |       |
|-------------------|-------|
| Best at TREC 2002 | 0.235 |
|-------------------|-------|

# Some Scores

Relevance scores (Avg F measure)

|                   |       |
|-------------------|-------|
| Human performance | 0.371 |
|-------------------|-------|

|        |       |
|--------|-------|
| Random | 0.040 |
|--------|-------|

|                   |       |
|-------------------|-------|
| Best at TREC 2002 | 0.235 |
|-------------------|-------|

|                 |       |
|-----------------|-------|
| UAmsStemmed (T) | 0.212 |
|-----------------|-------|

|                    |       |
|--------------------|-------|
| UAmsLemmatized (T) | 0.210 |
|--------------------|-------|

|                   |       |
|-------------------|-------|
| UAmsWordBased (T) | 0.205 |
|-------------------|-------|

|                           |       |
|---------------------------|-------|
| UAmsDocumentExpansion (T) | 0.199 |
|---------------------------|-------|

# More Findings

# More Findings

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.212         | 0.17     | 0.49    |
| UAmsStemmed (TD)  | 0.168         | 0.11     | 0.63    |
| UAmsStemmed (TDN) | 0.133         | 0.08     | 0.76    |

## More Findings

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.212         | 0.17     | 0.49    |
| UAmsStemmed (TD)  | 0.168         | 0.11     | 0.63    |
| UAmsStemmed (TDN) | 0.133         | 0.08     | 0.76    |

- ▶ No gains from linguistics clues

# More Findings

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.212         | 0.17     | 0.49    |
| UAmsStemmed (TD)  | 0.168         | 0.11     | 0.63    |
| UAmsStemmed (TDN) | 0.133         | 0.08     | 0.76    |

- ▶ No gains from linguistics clues
- ▶ Precision is the main challenge

# More Findings

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.212         | 0.17     | 0.49    |
| UAmsStemmed (TD)  | 0.168         | 0.11     | 0.63    |
| UAmsStemmed (TDN) | 0.133         | 0.08     | 0.76    |

- ▶ No gains from linguistics clues
- ▶ Precision is the main challenge
  - ongoing work: exploit the narrative

# More Findings

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.212         | 0.17     | 0.49    |
| UAmsStemmed (TD)  | 0.168         | 0.11     | 0.63    |
| UAmsStemmed (TDN) | 0.133         | 0.08     | 0.76    |

- ▶ No gains from linguistics clues
- ▶ Precision is the main challenge
  - ongoing work: exploit the narrative
  - ongoing work: develop more involved notions of relevance

# Novelty Part

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen
- ▶ Entailment  $\sim$  non-symmetric weighted overlap

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen
- ▶ Entailment  $\sim$  non-symmetric weighted overlap
- ▶ Which weights?

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen
- ▶ Entailment  $\sim$  non-symmetric weighted overlap
- ▶ Which weights? Inverse document frequency (*idf*)
  - $N$ : number of docs in sentences-as-docs collection
  - $n_i$ : number of docs in which the term  $t_i$  occurs

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen
- ▶ Entailment  $\sim$  non-symmetric weighted overlap
- ▶ Which weights? Inverse document frequency (*idf*)
  - $N$ : number of docs in sentences-as-docs collection
  - $n_i$ : number of docs in which the term  $t_i$  occurs

$$idf_i = \log \left( \frac{N}{n_i} \right)$$

# Novelty Part

- ▶ A sentence is **new** if it is not 'entailed' by what we have already seen
- ▶ Entailment  $\sim$  non-symmetric weighted overlap
- ▶ Which weights? Inverse document frequency (*idf*)
  - $N$ : number of docs in sentences-as-docs collection
  - $n_i$ : number of docs in which the term  $t_i$  occurs

$$idf_i = \log \left( \frac{N}{n_i} \right)$$

- terms  $t_i$  that occur in many docs have a high  $n_i$ , and hence a low *idf* score

# Novelty Part

# Novelty Part

- ▶ The **entailment score**  $es(s_i, s_j)$  of two (sets of) sentences

$$es(s_i, s_j) = \frac{\sum_{t_k \in (s_i \cap s_j)} idf_k}{\sum_{t_k \in s_j} idf_k}$$

# Novelty Part

- ▶ The **entailment score**  $es(s_i, s_j)$  of two (sets of) sentences

$$es(s_i, s_j) = \frac{\sum_{t_k \in (s_i \cap s_j)} idf_k}{\sum_{t_k \in s_j} idf_k}$$

- ‘how many of the content-bearing terms in  $s_j$  occur in  $s_i$ ?’

# Novelty Part

- ▶ The **entailment score**  $es(s_i, s_j)$  of two (sets of) sentences

$$es(s_i, s_j) = \frac{\sum_{t_k \in (s_i \cap s_j)} idf_k}{\sum_{t_k \in s_j} idf_k}$$

- ‘how many of the content-bearing terms in  $s_j$  occur in  $s_i$ ?’
- entailment threshold

# Novelty Part

- ▶ The **entailment score**  $es(s_i, s_j)$  of two (sets of) sentences

$$es(s_i, s_j) = \frac{\sum_{t_k \in (s_i \cap s_j)} idf_k}{\sum_{t_k \in s_j} idf_k}$$

- ‘how many of the content-bearing terms in  $s_j$  occur in  $s_i$ ?’
- entailment threshold
- ▶ Go down list of relevant sentences, taking first sentence as starting point and include later ones only if not entailed by ones already included

# Some Novelty Scores

# Some Novelty Scores

|                   | Novelty scores (Avg F measure) |
|-------------------|--------------------------------|
| Human performance | 0.353                          |
| Random            | 0.037                          |
| Best at TREC 2002 | 0.217                          |
| UAmSStemmed (T)   | 0.205                          |

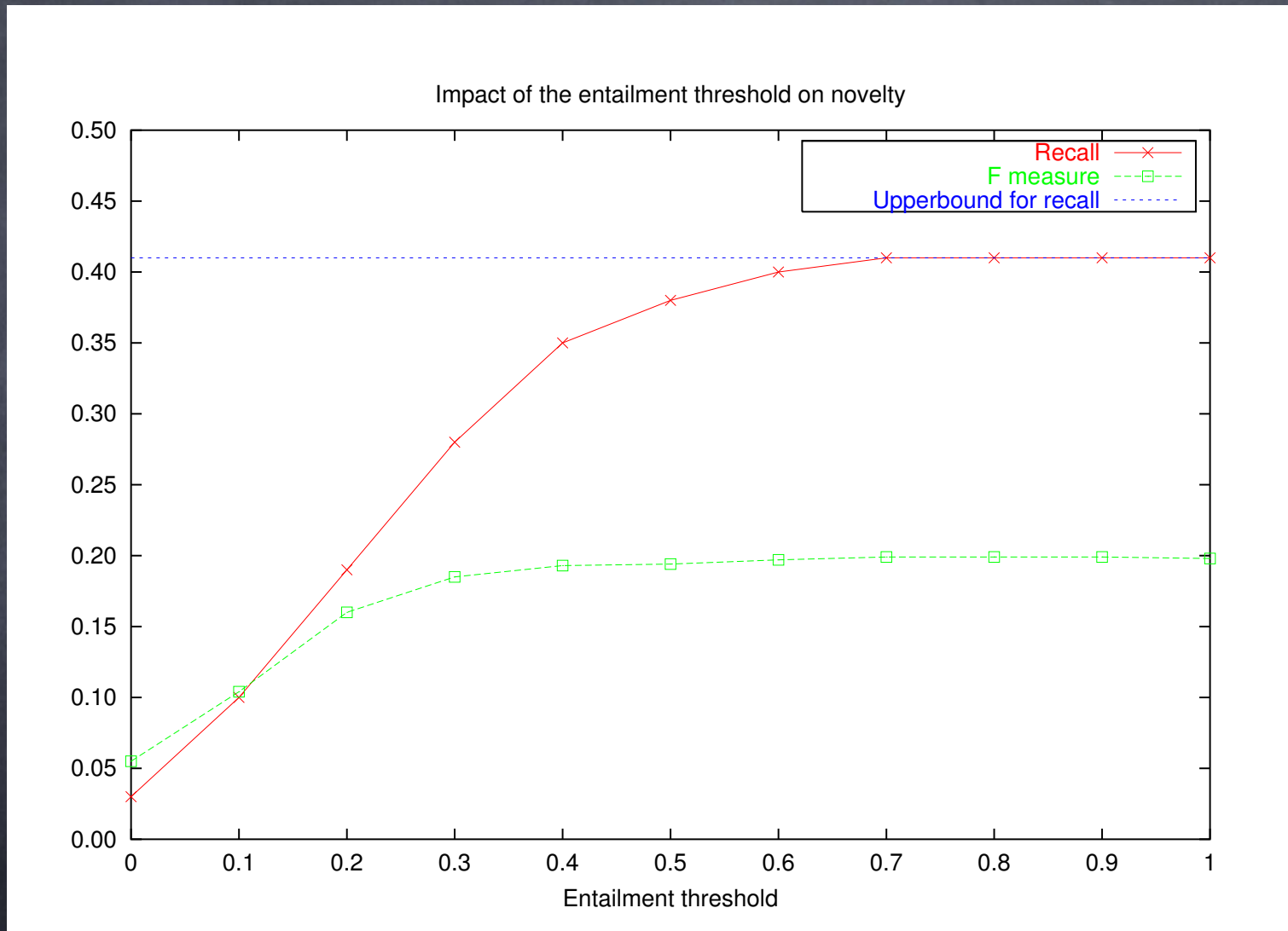
# Some Novelty Scores

| Novelty scores (Avg F measure) |       |
|--------------------------------|-------|
| Human performance              | 0.353 |
| Random                         | 0.037 |
| Best at TREC 2002              | 0.217 |
| UAmsStemmed (T)                | 0.205 |

|                   | Avg F measure | Avg Prec | Avg Rec |
|-------------------|---------------|----------|---------|
| UAmsStemmed (T)   | 0.205         | 0.16     | 0.48    |
| UAmsLemm. (T)     | 0.202         | 0.16     | 0.50    |
| UAmsWordBased (T) | 0.199         | 0.18     | 0.41    |
| UAmsDocExp (T)    | 0.182         | 0.13     | 0.51    |

# How Good can It Get?

# How Good can It Get?



# How Much Inference?

# How Much Inference?

- ▶ Entailment score is naive. . . what is the potential for more sophisticated notions of inference?
- ▶ Need richer representations of docs and topics
  - named entity extraction
  - relation extraction through dependency parsing

# How Much Inference?

- ▶ Entailment score is naive. . . what is the potential for more sophisticated notions of inference?
- ▶ Need richer representations of docs and topics
  - named entity extraction
  - relation extraction through dependency parsing
- ▶ Results (based on experiments by other participants in the TREC novelty task, and own informal ones)
  - less robust
  - poor scores for relevance part

# How Much Inference?

- ▶ Entailment score is naive. . . what is the potential for more sophisticated notions of inference?
- ▶ Need richer representations of docs and topics
  - named entity extraction
  - relation extraction through dependency parsing
- ▶ Results (based on experiments by other participants in the TREC novelty task, and own informal ones)
  - less robust
  - poor scores for relevance part
- ▶ Anyway: simple entailment scoring seems good enough

# Novelty: Summarizing

# Novelty: Summarizing

- ▶ One of several tasks aimed at pinpointing relevant information, allowing for easy experimentation

# Novelty: Summarizing

- ▶ One of several tasks aimed at pinpointing relevant information, allowing for easy experimentation
- ▶ Main challenge is to increase precision

# Novelty: Summarizing

- ▶ One of several tasks aimed at pinpointing relevant information, allowing for easy experimentation
- ▶ Main challenge is to increase precision
- ▶ Ongoing work is aimed at exploiting negative information in the narrative and at exploring alternative notions of context

# Novelty: Summarizing

- ▶ One of several tasks aimed at pinpointing relevant information, allowing for easy experimentation
- ▶ Main challenge is to increase precision
- ▶ Ongoing work is aimed at exploiting negative information in the narrative and at exploring alternative notions of context
- ▶ Limited potential for deep inference

# Extracting Opinions

# Extracting Opinions

- ▶ Large parts of human communication is not about matter-of-fact content, but about **subjective assessments** and **opinions**

# Extracting Opinions

- ▶ Large parts of human communication is not about matter-of-fact content, but about **subjective assessments** and **opinions**
- ▶ How can we extract subjective meaning (preferably from largish bodies of text) in a **robust** manner?

# Extracting Opinions

- ▶ Large parts of human communication is not about matter-of-fact content, but about **subjective assessments** and **opinions**
- ▶ How can we extract subjective meaning (preferably from largish bodies of text) in a **robust** manner?
- ▶ Approach: use lexical resources to 'map out' opinions

# Extracting Opinions

- ▶ Large parts of human communication is not about matter-of-fact content, but about **subjective assessments** and **opinions**
- ▶ How can we extract subjective meaning (preferably from largish bodies of text) in a **robust** manner?
- ▶ Approach: use lexical resources to ‘map out’ opinions
- ▶ Test scenarios: Internet discussion groups on elections, companies, politics, music, . . .

# WordNet

# WordNet

- ▶ WordNet is a big lexical database in which words are connected by synonymy
  - gives rise to a natural distance between two words
  - use this distance metric to assess come to grips with the subjective meaning of words

# WordNet

- ▶ WordNet is a big lexical database in which words are connected by synonymy
  - gives rise to a natural distance between two words
  - use this distance metric to assess come to grips with the subjective meaning of words
- ▶ WordNet has several giant components: one has adjectives that express **affective** or **emotive** meaning
  - 5410 adjective words (is 25.3%)
  - 5464 adjective synsets (is 29.5%)
  - diameter (maximal distance) is 26, mean dist is 9.25
  - descriptive adjectives like *good* (*bad*) and *beautiful* (*ugly*)

# Osgood

# Osgood

- ▶ Charles Osgood, in the 1950s, did research on affective meaning (*The Meaning of Meaning*)

# Osgood

- ▶ Charles Osgood, in the 1950s, did research on affective meaning (*The Meaning of Meaning*)
- ▶ Factor analysis revealed three dominant factors of meaning
  - **evaluative** factor (*good–bad*)
  - **potency** factor (*strong–weak*)
  - **activity** factor (*active–passive*)

# Osgood

- ▶ Charles Osgood, in the 1950s, did research on affective meaning (*The Meaning of Meaning*)
- ▶ Factor analysis revealed three dominant factors of meaning
  - **evaluative** factor (*good–bad*)
  - **potency** factor (*strong–weak*)
  - **activity** factor (*active–passive*)
- ▶ All of Osgood's bipolar adjectives are in the giant component

# Using the Structure of WordNet

# Using the Structure of WordNet

- ▶ Use distance to the adjective *good* as indicator for goodness?

# Using the Structure of WordNet

- ▶ Use distance to the adjective *good* as indicator for goodness?
- ▶ The answer is **NO!**
  - distance between antonyms *good* and *bad* is 4 (mean is 9.25!)

# Using the Structure of WordNet

- ▶ Use distance to the adjective *good* as indicator for goodness?
- ▶ The answer is **NO!**
  - distance between antonyms *good* and *bad* is 4 (mean is 9.25!)
- ▶ For each word in the giant component, we have distances to both *good* and *bad*.
  - what if we consider both distances?

# Using the Structure of WordNet

- ▶ Use distance to the adjective *good* as indicator for goodness?
- ▶ The answer is **NO!**
  - distance between antonyms *good* and *bad* is 4 (mean is 9.25!)
- ▶ For each word in the giant component, we have distances to both *good* and *bad*.
  - what if we consider both distances?
- ▶ Watching WordNet

# Measure for Distance to 'good' and 'bad'

# Measure for Distance to 'good' and 'bad'

- ▶ For the evaluative factor, calculate

$$\text{EVA}(word) = \frac{\text{distance}(word, \text{bad}) - \text{distance}(word, \text{good})}{\text{distance}(\text{good}, \text{bad})}$$

- this gives a value in  $[-1, 1]$
- use this (partial) function to rate texts on this dimension

# Measure for Distance to 'good' and 'bad'

- ▶ For the evaluative factor, calculate

$$\text{EVA}(word) = \frac{\text{distance}(word, \text{bad}) - \text{distance}(word, \text{good})}{\text{distance}(\text{good}, \text{bad})}$$

- this gives a value in  $[-1, 1]$
  - use this (partial) function to rate texts on this dimension
- ▶ Similar for other Osgood factors:
    - **potency** factor with distances to **strong** and **weak**
    - **activity** factor with distances to **active** and **passive**

# Implementation

# Implementation

- ▶ Expensive to determine shortest distances
  - use a precompiled list of all 5410 adjectives
  - unrelated words score 0

# Implementation

- ▶ Expensive to determine shortest distances
  - use a precompiled list of all 5410 adjectives
  - unrelated words score 0
- ▶ Is the list biased?
  - 35 synonyms of the adjective 'good'
  - and only 15 synonyms of 'bad'

# Implementation

- ▶ Expensive to determine shortest distances
  - use a precompiled list of all 5410 adjectives
  - unrelated words score 0
- ▶ Is the list biased?
  - 35 synonyms of the adjective 'good'
  - and only 15 synonyms of 'bad'
- ▶ **No:** it is surprisingly balanced
  - all 5410 adjective get a value from  $-1$  to  $1$
  - total sum of values is  $-48.25$
  - mean value of  $\frac{-48.25}{5410} = -0.0089$

# Evaluation

# Evaluation

- ▶ The **General Inquirer** (Stone 1966) is the classic system for content analysis

# Evaluation

- ▶ The **General Inquirer** (Stone 1966) is the classic system for content analysis
- ▶ We can evaluate our WordNet measures against the manually constructed lists of the General Inquirer:

# Evaluation

- ▶ The **General Inquirer** (Stone 1966) is the classic system for content analysis
- ▶ We can evaluate our WordNet measures against the manually constructed lists of the General Inquirer:

| Factor     | Common words | Correct |                       |
|------------|--------------|---------|-----------------------|
| Evaluative | 317          | 68.19%  |                       |
| Potency    | 365          | 71.36%  |                       |
| Activity   | 115          | 61.71%  |                       |
| Evaluative | 618          | 67.64%  | (new enlarged GI-set) |

# Internet Opinion Polling

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .
- ▶ Outline of prototype analyzing newsgroup submissions
  - (1) For a given **subject**, we identify relevant newsgroups

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .
- ▶ Outline of prototype analyzing newsgroup submissions
  - (1) For a given **subject**, we identify relevant newsgroups
  - (2) Each day, we collect all headings from postings, and match the subject-line against the **subject**

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .
- ▶ Outline of prototype analyzing newsgroup submissions
  - (1) For a given **subject**, we identify relevant newsgroups
  - (2) Each day, we collect all headings from postings, and match the subject-line against the **subject**
  - (3) If it matches, we collect the bodies of these submission

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .
- ▶ Outline of prototype analyzing newsgroup submissions
  - (1) For a given **subject**, we identify relevant newsgroups
  - (2) Each day, we collect all headings from postings, and match the subject-line against the **subject**
  - (3) If it matches, we collect the bodies of these submission
  - (4) We filter for junk, and rate the used words on Osgood's dimensions

# Internet Opinion Polling

- ▶ Lifting method from individual words to documents. . .
- ▶ Outline of prototype analyzing newsgroup submissions
  - (1) For a given **subject**, we identify relevant newsgroups
  - (2) Each day, we collect all headings from postings, and match the subject-line against the **subject**
  - (3) If it matches, we collect the bodies of these submission
  - (4) We filter for junk, and rate the used words on Osgood's dimensions
- ▶ Some examples. . .

# Test Domains for Opinion Extraction

# Test Domains for Opinion Extraction

- ▶ Teenage pop stars

# Test Domains for Opinion Extraction

- ▶ Teenage pop stars
- ▶ Fortune 500 companies

# Test Domains for Opinion Extraction

- ▶ Teenage pop stars
- ▶ Fortune 500 companies
- ▶ UK Politics
  - tracking since the 2001 parliamentary election campaign

# Test Domains for Opinion Extraction

- ▶ Teenage pop stars
- ▶ Fortune 500 companies
- ▶ UK Politics
  - tracking since the 2001 parliamentary election campaign
- ▶ German politics

# Test Domains for Opinion Extraction

- ▶ Teenage pop stars
- ▶ Fortune 500 companies
- ▶ UK Politics
  - tracking since the 2001 parliamentary election campaign
- ▶ German politics
- ▶ Dutch politics
  - recently started tracking newsgroups on Dutch politics
  - received extensive coverage in *NRC Handelsblad* (~ *La Repubblica*)

# Background on Dutch Politics

# Background on Dutch Politics

## ▶ 2001

- wide dissatisfaction with traditional political parties, especially those in government
- Lijst Pim Fortuyn (LPF, named after its leader) is founded and gains instant popularity

# Background on Dutch Politics

## ▶ 2001

- wide dissatisfaction with traditional political parties, especially those in government
- Lijst Pim Fortuyn (LPF, named after its leader) is founded and gains instant popularity

## ▶ 2002

- May 6: leader of the LPF is murdered
- May 13: elections, LPF gains 17% of the votes (up from 0%!)
- July 22: new government installed, a coalition of 3 parties including LPF

# More Background on Dutch Politics

# More Background on Dutch Politics

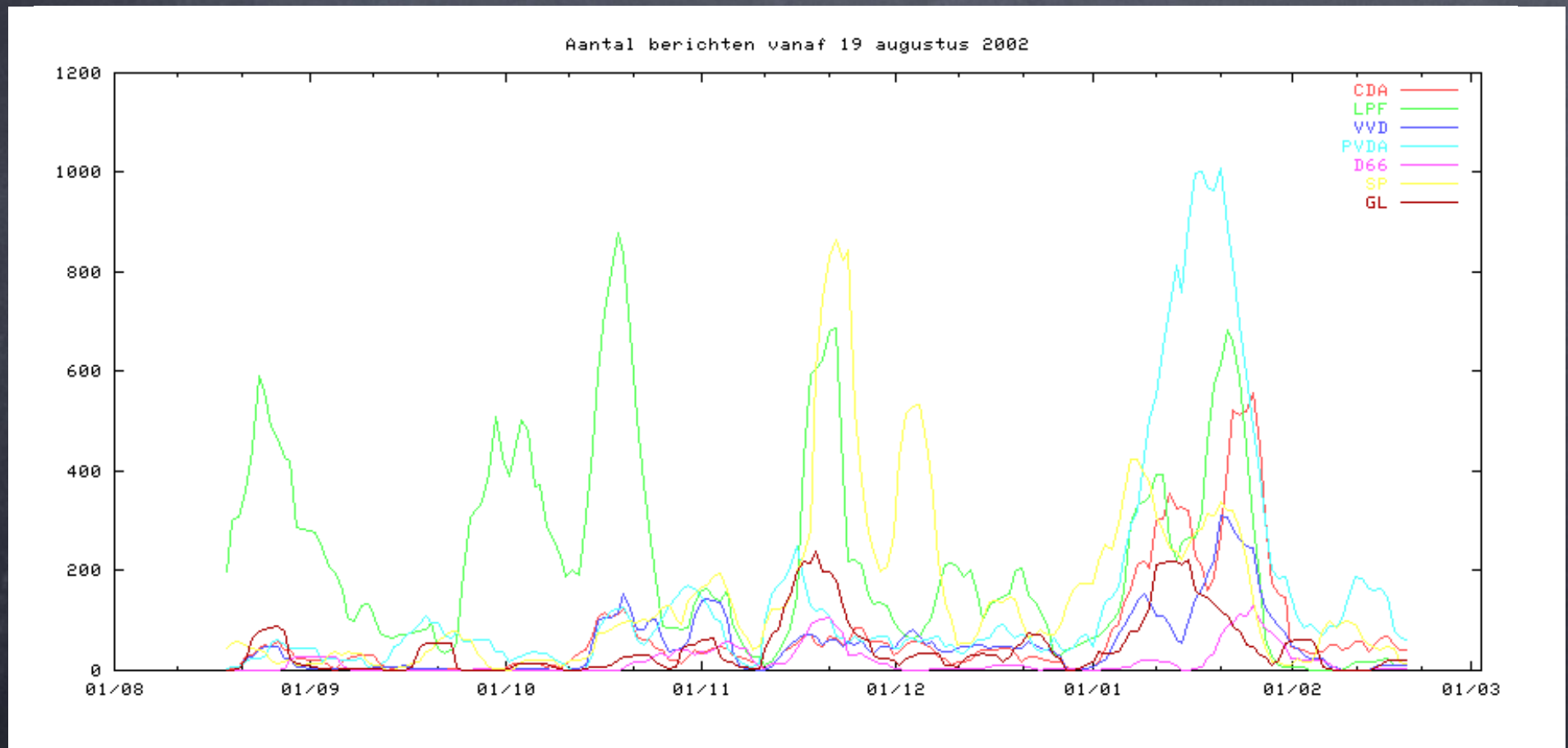
- ▶ 2002 (continued)
  - October 16: government falls; LPF at center of problems; shortest serving government since WWII (86 days)

# More Background on Dutch Politics

- ▶ 2002 (continued)
  - October 16: government falls; LPF at center of problems; shortest serving government since WWII (86 days)
- ▶ 2003
  - January 22: elections, LPF down from 17% to 5%
  - 'old' order restored, with the traditional parties very strongly back in the polls
  - February 25: still no new government

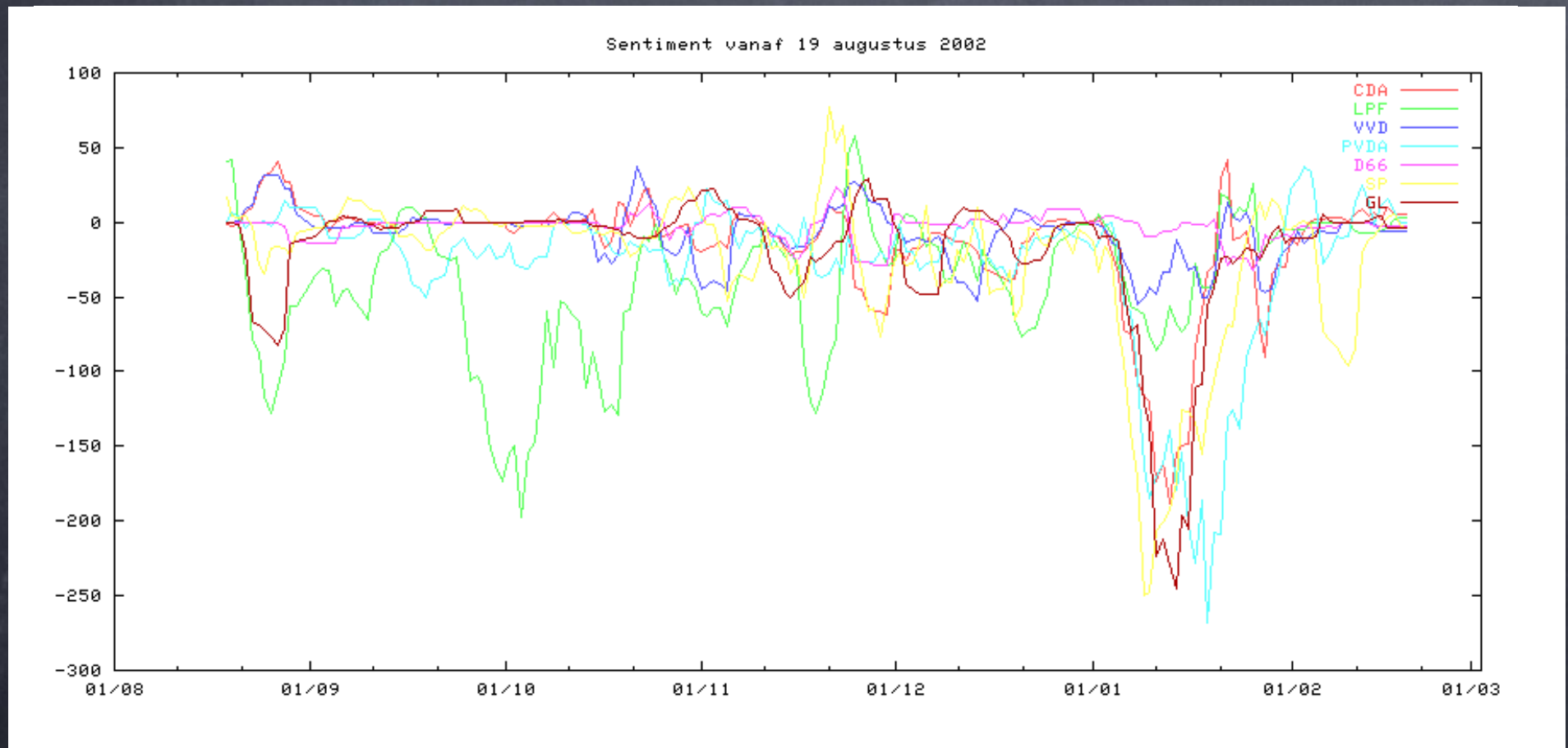
# Dutch Politics since August 19, 2002

# Dutch Politics since August 19, 2002



# Dutch Politics since August 19, 2002

# Dutch Politics since August 19, 2002



# Opinion Extraction: Summarizing

# Opinion Extraction: Summarizing

- ▶ Opinions matter. . .

# Opinion Extraction: Summarizing

- ▶ Opinions matter. . .
- ▶ Financial Times, Feb 23, 2003: 'Google buys Pyra Labs' (developers of Blogger)

*. . . . Marry up all that user-generated content with a powerful search engine, and a wealth of information and **opinion** could become more accessible to a general web audience . . . .*

# Opinion Extraction: Summarizing

# Opinion Extraction: Summarizing

- ▶ Opinionated content proliferates on the Internet
  - Internet news sites
  - Internet discussion sites

# Opinion Extraction: Summarizing

- ▶ Opinionated content proliferates on the Internet
  - Internet news sites
  - Internet discussion sites
- ▶ Tools for extracting subjective meaning turn the Internet into an automatic opinion polling system
  - at virtually no cost, on virtually any topic (politics, stocks, books, music, . . . ), with daily updates, even near-real time for active topics

# Opinion Extraction: Summarizing

- ▶ Opinionated content proliferates on the Internet
  - Internet news sites
  - Internet discussion sites
- ▶ Tools for extracting subjective meaning turn the Internet into an automatic opinion polling system
  - at virtually no cost, on virtually any topic (politics, stocks, books, music, . . . ), with daily updates, even near-real time for active topics
- ▶ Still much to do. . .
  - working on more detailed measures for subjective meaning
  - working on validation and evaluation at doc level

# Wrap-Up

# Wrap-Up

- ✓ Beyond document retrieval
  - novelty track
  - opinion extraction

# Wrap-Up

- ✓ Beyond document retrieval
  - novelty track
  - opinion extraction
- ✓ Intelligent information access requires a mixture of information retrieval, natural language processing, and artificial intelligence

# Wrap-Up

- ✓ Beyond document retrieval
  - novelty track
  - opinion extraction
- ✓ Intelligent information access requires a mixture of information retrieval, natural language processing, and artificial intelligence

Language and Inference Technology Group

<http://lit.science.uva.nl>