Creating Specialized and General Corpora Using Automated Search Engine Queries

Marco Baroni, Serge Sharoff

SSLMIT, University of Bologna; CTS, University of Leeds

Birmingham, July 2005

Outline

1. Introduction
2. Building a specialized corpus
   - Background
   - The procedure in detail
   - Conclusions on specialized corpus building
3. Building a BNC
   - DIY manual
   - Analysing macrostructure (composition)
   - Analysing microstructure (lexicon)
4. Conclusions

Introduction

- A “middle ground strategy”

Introduction

- A “middle ground strategy”
- Some relevant work:
  - Ghani and colleagues’ CorpusBuilder project
  - Corpus comparison work, e.g., Rayson and Garside 2000

What you need

- Unix-like OS and Unix skills

What you need

- Unix-like OS and Unix skills
- Google API (or, now, Yahoo API)

What you need

- Unix-like OS and Unix skills
- Google API (or, now, Yahoo API)
- Our scripts – contact us!

What you need

- Unix-like OS and Unix skills
- Google API (or, now, Yahoo API)
- Our scripts – contact us!
- POS taggers, indexers, etc.

Outline

1. Introduction
2. Building a specialized corpus
   - Background
   - The procedure in detail
   - Conclusions on specialized corpus building
3. Building a BNC
   - DIY manual
   - Analysing macrostructure (composition)
   - Analysing microstructure (lexicon)
4. Conclusions

Applications

- Uses: technical translation, terminography, populating ontologies. . .

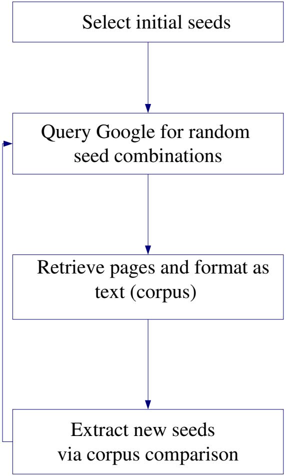
Applications

- Uses: technical translation, terminography, populating ontologies. . .
- Domains: medical, legal, meteorology, arts, food, nautical terminology, (e-)commerce. . .

Applications

- Uses: technical translation, terminography, populating ontologies. . .
- Domains: medical, legal, meteorology, arts, food, nautical terminology, (e-)commerce. . .
- Languages: English, Italian, Japanese, Spanish, German, French, Danish. . .

The basic idea



Select initial seeds

Query Google for random seed combinations

Retrieve pages and format as text (corpus)

Extract new seeds via corpus comparison

Example

- 7 seeds: *black sabbath*, *led zeppelin*, *deep purple*, *motorhead*, *rainbow*, *judas priest*, *iron maiden*

Example

- 7 seeds: *black sabbath*, *led zeppelin*, *deep purple*, *motorhead*, *rainbow*, *judas priest*, *iron maiden*
- 35 3-seed combinations:
  "led zeppelin" rainbow "black sabbath"
  "deep purple" motorhead rainbow
  "deep purple" "judas priest" motorhead
  ...

Example

- 7 seeds: *black sabbath*, *led zeppelin*, *deep purple*, *motorhead*, *rainbow*, *judas priest*, *iron maiden*
- 35 3-seed combinations:
  "led zeppelin" rainbow "black sabbath"
  "deep purple" motorhead rainbow
  "deep purple" "judas priest" motorhead
  ...
- 20 documents per query

Document retrieval and processing

- Automated retrieval of documents is the easy part (e.g., with perl LWP module). . .

Document retrieval and processing

- Automated retrieval of documents is the easy part (e.g., with perl LWP module). . .
- Filtering and cleaning (“boilerplate removal”) is more tricky

Cleaning with a standard HTML formatter

```
Blackmore’s Night Latest News
Ritchie Blackmore’s Bio
Blackmore’s Night Band Bios
Blackmore’s Night Tour Info
Blackmore’s Night Merchandise
Blackmore’s Night Photo Gallery
Blackmore’s Night Audio Clips
...
Register for
Blackmores Night
Email Updates!
Just enter your
email address in
the box below and
click the ’Sign up’ button!
...
RITCHIE BLACKMORE A MUSICAL HISTORY...
1967 - RITCHIE BLACKMORE - who has previously played with such bands
as the Outlaws, Screaming Lord Sutch, and Neil Christian & The
Crusaders - is invited by ex-Artwoods/The Flowerpot Men keybordist Jon
Lord (who was invited by The Searchers ex-drummer, Chris Curtis) to
form a new band. Other musician’s would be auditioned from a Melody
Maker ad in Deeves Hall in Hertfordshire.
1968- In February, the group would form as Roundabout, consisting of
the three (with Chris Curtis on vocals) along with Dave Curtis on bass
and Bobby Woodman on drums. After only a month of uncompromising
rehearsals, BLACKMORE and LORD would be the only two remaining,
...
```

Finn’s BTE heuristic

- `http://www.smi.ucd.ie/hyppia/`

Finn’s BTE heuristic

- `http://www.smi.ucd.ie/hyppia/`
- Basic observation: Content-rich section of page tends to occur in low-HTML-density area

Finn’s BTE heuristic

- `http://www.smi.ucd.ie/hyppia/`
- Basic observation: Content-rich section of page tends to occur in low-HTML-density area
- Look for stretch that maximizes the quantity: $N(TOKEN) - N(TAG)$

Finn’s BTE heuristic: why it (mostly) works

```
TAG TAG TOKEN TOKEN TAG TAG TAG
TOKEN TAG TAG
TOKEN TAG TAG
TAG TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TOKEN TAG TOKEN TOKEN TAG TOKEN TOKEN TOKEN
TAG TAG
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TAG TAG TAG TAG TAG
TAG TOKEN TAG TAG TOKEN TAG
```

Cleaning with Finn’s heuristic

```
- RITCHIE BLACKMORE - who has previously played with such bands as the Outlaws, Screaming Lor
d Sutch, and Neil Christian & The Crusaders - is invited by ex-Artwoods/The Flowerpot Men key
bordist Jon Lord (who was invited by The Searchers ex-drummer, Chris Curtis) to form a new ba
nd. Other musician’s would be auditioned from a Melody Maker ad in Deeves Hall in Hertfordshi
re. 1968 - In February, the group would form as Roundabout, consisting of the three (with Chr
is Curtis on vocals) along with Dave Curtis on bass and Bobby Woodman on drums. After only a 
month of uncompromising rehearsals, BLACKMORE and LORD would be the only two remaining, bring
ing in vocalist Rod Evans (formally of the groups M15 and Maze), bassist Nick Simper (ex-John
ny Kidd & the Pirates) and drummer Ian Paice. In April, the band would change its name to DEE
P PURPLE, and forego the name Concrete God. Using Vanilla Fudge as its model, the group recor
ds an album and is signed to EMI in the United Kingdom and Tetragrammation (Bill Cosby’s labe
l) in the United States. A few months later, the band performs its first major UK performance
 at the Sunbury Festival. In September, the band would release the first single from the albu
m, "Hush", which reaches #4 on the US charts. The album, itself, Shades of Deep Purple, would
 reach the Top 25. In December, the band would release their rendition of Neil Diamond’s "Ken
tucky Woman" a single which would also experience considerable chart success. 1969 - Early in
 the year, the band would release The Book to Taliesyn which would feature a revival of Ike a
nd Tina Turner’s "River Deep, Mountain High." In July, the band would release their self-titl
ed album, Deep Purple. Shortly thereafter, their record label would go under and Rod Evans an
d Nick Simper would both leave the band. They would be replaced by singer Ian Gillan and bass
ist Roger Glover, who were both previously members of the group, Episode Six. The two would a
dd a dynamic chemistry to the band, complimenting BLACKMORE, who was quickly rising as one of
 most omnipotent British lead guitarists around. In September, the band would perform what ma
ny believed to be their most adventurous album to date, Concerto For Group And Orchestra, whi
ch featured The Royal Philharmonic Orchestra at the Royal Albert Hall in London. 1970 - By Au
gust, the band’s newest album, Deep Purple In Rock, would stay on the UK charts for over a ye
ar, and hit #4. To this day, the album remains a landmark of sorts in heavy rock, with Gillan
...
```

Further filtering

- Pre-HMTL-Stripping: Minimum and maximum size, e.g., 5-200 Kb (Fletcher's suggestion, confirmed by my experiments)

Further filtering

- Pre-HMTL-Stripping: Minimum and maximum size, e.g., 5-200 Kb (Fletcher's suggestion, confirmed by my experiments)
- Post-Boilerplate-Removal: Minimum number and proportion of function/frequent words (Zipf's law to the rescue!)

Example continued: size of retrieved corpus

- Raw: 12MB
- Removing HTML etc.: 2.8MB
- After boilerplate-removal and Zipfian filtering: 1.3MB, 281 documents, ∼217K words

New seed extraction

- Extract typical terms through statistical comparison with reference corpus (using Log-Likelihood Ratio, Mutual Information, etc.)

New seed extraction

- Extract typical terms through statistical comparison with reference corpus (using Log-Likelihood Ratio, Mutual Information, etc.)
- Use extracted terms as new seeds to build a larger corpus

New seed extraction

- Extract typical terms through statistical comparison with reference corpus (using Log-Likelihood Ratio, Mutual Information, etc.)
- Use extracted terms as new seeds to build a larger corpus
- Reference corpus: you can build your own from the Web (see below)!

New seed extraction

- Extract typical terms through statistical comparison with reference corpus (using Log-Likelihood Ratio, Mutual Information, etc.)
- Use extracted terms as new seeds to build a larger corpus
- Reference corpus: you can build your own from the Web (see below)!
- For all your word statistics needs, there is the UCS package:
  `http://www.collocations.de`

Example continued

Top 100 words found with Log-Likelihood Ratio (using the Brown as a reference corpus) as new seeds, for example:

band
album
metal
rock
dio
tour
live
iommi
released
heavy
ozzy
osbourne
bands
music
release
guitar
song
drummer
albums
blackmore

Example continued

Some words that would have been in seeds if we had not performed boilerplate-removal / Zipfian filtering:

alice
avg
bestel
click
floyd
hendrix
min
modblog
news
pantera
picture
reviews
slayer
t-shirts

Example continued: the second round

- 200 5-seed combinations used as queries

Example continued: the second round

- 200 5-seed combinations used as queries
- 1150 pages retrieved

Example continued: the second round

- 200 5-seed combinations used as queries
- 1150 pages retrieved
- ∼3.6M words (after filtering)

Example continued: the second round

- 200 5-seed combinations used as queries
- 1150 pages retrieved
- ∼3.6M words (after filtering)
- Top 20 words by Log-Likelihood Ratio:

band
album
rock
sabbath
metal
music
guitar
tour
song
black
songs
ozzy
live
released
release
bands
purple
vocals
bass
albums

Issues and directions

- “Know-how” (how many seeds? combinations? iterations?)

Issues and directions

- “Know-how” (how many seeds? combinations? iterations?)
- What is a specialized corpus? Which kinds of specialization “work”/”do not work”?

Issues and directions

- “Know-how” (how many seeds? combinations? iterations?)
- What is a specialized corpus? Which kinds of specialization “work”/”do not work”?
- Other kinds of specialization: study of genre, personal prose, everyday domains, child-directed prose, etc.

Outline

1. Introduction
2. Building a specialized corpus
   - Background
   - The procedure in detail
   - Conclusions on specialized corpus building
3. Building a BNC
   - DIY manual
   - Analysing macrostructure (composition)
   - Analysing microstructure (lexicon)
4. Conclusions

Biting the bullet. . .

- Crawling, cleaning, annotating, managing and maintaining your own indexed version of the web.
- Obviously, the “ideal” solution.
- But obviously a lot of work!
- A shortcut: building a BNC for language X using BootCat

Step 1: Word selection

- Choose 500 word forms frequent in X

Step 1: Word selection

- Choose 500 word forms frequent in X
- If X=English: *events, picture*, . . .

Step 1: Word selection

- Choose 500 word forms frequent in X
- If X=English: *events, picture*, . . .
- Not function words, not specific words

Step 1: Word selection

- Choose 500 word forms frequent in X
- If X=English: *events, picture*, . . .
- Not function words, not specific words
- If X=Russian: события, картина, . . .

Step 1: Word selection

- Choose 500 word forms frequent in X
- If X=English: *events, picture*, . . .
- Not function words, not specific words
- If X=Russian: события, картина, . . .
- Problems with morphology: посылать - 64 forms

Step 1: Word selection

- Choose 500 word forms frequent in X
- If X=English: *events, picture*, . . .
- Not function words, not specific words
- If X=Russian: события, картина, . . .
- Problems with morphology: посылать - 64 forms
- But the advantage of morphologically rich languages: we can use verbs only to get descriptive fragments

Step 2: Query generation

- Build 5,000 queries 4 words each and send them to Google

Step 2: Query generation

- Build 5,000 queries 4 words each and send them to Google
- Restrict the output to language X, HTML only, allintext

Step 2: Query generation

- Build 5,000 queries 4 words each and send them to Google
- Restrict the output to language X, HTML only, allintext
- If X is not listed (Ukrainian), add a frequent function word (має OR її)

Step 2: Query generation

- Build 5,000 queries 4 words each and send them to Google
- Restrict the output to language X, HTML only, allintext
- If X is not listed (Ukrainian), add a frequent function word (має OR її)
- Results for *picture* AND *extent* AND *raised* AND *events*: `http://www.google.com/search?q=picture+extent+raised+events&lr=lang_en`

Step 2: Query generation

- Build 5,000 queries 4 words each and send them to Google
- Restrict the output to language X, HTML only, allintext
- If X is not listed (Ukrainian), add a frequent function word (має OR її)
- Results for *picture* AND *extent* AND *raised* AND *events*: `http://www.google.com/search?q=picture+extent+raised+events&lr=lang_en`
- All retrieved examples contain connected text (2000-5000 words)

Step 3: Downloading

- Maintain the list of top 10 URLs for each query
- This is an open source corpus—
  We are free to distribute the list of URLs
- Download available pages (direct or from the Google cache)
- Currently we have I-DE, I-EN, I-RU, I-ZH (+I-RO, I-UA)

Step 4: Post-processing

- Unify page encoding (6 encodings for Russian)
- Convert pages to plain text (using Lynx)
- Remove navigation frames/boilerplates
- Filter out duplicates and near duplicates
- Language-dependent morphosyntactic processing (tokenisation, lemmatisation, POS tagging, parsing, word sense disambiguation, . . .)
- Indexing in CWB

Validation

- We have 45,000 documents—
  a random snapshot of the Internet for language X
- How random is the procedure?
- What is there: benchmarking the macro- and microstructure

Data for comparing text composition

- Use a text typology (a set of functional categories)

Data for comparing text composition

- Use a text typology (a set of functional categories)
- Design detection criteria

Data for comparing text composition

- Use a text typology (a set of functional categories)
- Design detection criteria
- Take a sample and code it using O'Donnel's Systemic Coder `http://www.wagsoft.com/Coder/`

Data for comparing text composition

- Use a text typology (a set of functional categories)
- Design detection criteria
- Take a sample and code it using O'Donnel's Systemic Coder `http://www.wagsoft.com/Coder/`
- Symmetric confidence interval: $\sigma = c\sqrt{\frac{p(1-p)}{N}}$

Data for comparing text composition

- Use a text typology (a set of functional categories)
- Design detection criteria
- Take a sample and code it using O'Donnel's Systemic Coder `http://www.wagsoft.com/Coder/`
- Symmetric confidence interval: $\sigma = c\sqrt{\frac{p(1-p)}{N}}$
- 200 documents – $\sigma = \pm 5\%, 90\%$ confidence (c=1.645) OR

Data for comparing text composition

- Use a text typology (a set of functional categories)
- Design detection criteria
- Take a sample and code it using O'Donnel's Systemic Coder `http://www.wagsoft.com/Coder/`
- Symmetric confidence interval: $\sigma = c\sqrt{\frac{p(1-p)}{N}}$
- 200 documents – $\sigma = \pm 5\%$, 90% confidence (c=1.645) OR
- 1500 documents – $\sigma = \pm 1\%$, 99% confidence (c=1.946)

Results: I-EN vs. BNC

- **Authorship**: **corporate: 44% (22%)**, male: 23% (28%), **female:4%(13%)**,unknown:7%(0%),multiple:19%(23%)

Results: I-EN vs. BNC

- **Authorship**: **corporate: 44% (22%)**, male: 23% (28%), **female:4%(13%)**,unknown:7%(0%),multiple:19%(23%)
- **Mode**: written:86%(90%), **electronic: 13%(0%)**, spoken:1%(10%)

Results: I-EN vs. BNC

- **Authorship**: **corporate: 44% (22%)**, male: 23% (28%), **female:4%(13%)**,unknown:7%(0%),multiple:19%(23%)
- **Mode**: written:86%(90%), **electronic: 13%(0%)**, spoken:1%(10%)
- **Audience**: general: 33%(27%), informed: 45%(47%), professional: 22%(26%)

Results: I-EN vs. BNC

- **Authorship**: **corporate: 44% (22%)**, male: 23% (28%), **female:4%(13%)**,unknown:7%(0%),multiple:19%(23%)
- **Mode**: written:86%(90%), **electronic: 13%(0%)**, spoken:1%(10%)
- **Audience**: general: 33%(27%), informed: 45%(47%), professional: 22%(26%)
- **Aims**: discussion: 45%, recommendation: 34%, information: 11%, instruction: 6%, **recreation: 4%**

Results: I-EN vs. BNC

- **Authorship**: **corporate: 44% (22%)**, male: 23% (28%), **female:4%(13%)**,unknown:7%(0%),multiple:19%(23%)
- **Mode**: written:86%(90%), **electronic: 13%(0%)**, spoken:1%(10%)
- **Audience**: general: 33%(27%), informed: 45%(47%), professional: 22%(26%)
- **Aims**: discussion: 45%, recommendation: 34%, information: 11%, instruction: 6%, **recreation: 4%**
- **Domain**: life: 14%(27%), politics: 12%(19%), commerce: 13%(8%), natsci: 3%(4%), **appsci: 29%(7%)**, socsci: 16%(17%), arts: 2%(7%), leisure: 11%(11%)

I-EN vs. BNC

More in BNC	LL-score	More in I-EN	LL-score
was	1251.29	your	303.43
had	953.62	Posted	278.37
he	928.66	Web	262.23
she	912.82	program	255.15
er	909.30	Internet	228.45
her	795.37	site	217.36
Yeah	623.65	Click	201.91
it	580.80	Center	192.76
erm	578.10	online	189.36
his	496.03	Bush	177.53
I	415.54	email	177.42
said	398.64	information	174.04
Oh	385.29	New	168.38

BNC vs. Reuters; I-EN vs. Reuters

More in BNC	LL-score	More in Reuters	LL-score
you	6005.14	say	8559.54
I	5271.42	percent	4513.35
she	3334.57	million	2364.29
be	2411.89	market	1982.47
do	1610.71	billion	1518.25
they	1502.79	bank	1468.84
your	1282.15	company	1258.34
can	1191.74	newsroom	1240.37
what	1090.53	share	1214.84
my	1023.56	tuesday	1199.25

More in I-EN	LL-score	More in Reuters	LL-score
you	4343.16	say	12154.94
I	2797.67	percent	3424.40
your	2731.17	million	2103.23
or	1845.60	market	1943.17
my	1262.80	bank	1574.68
can	965.08	billion	1270.30
this	899.29	newsroom	1254.03
use	729.11	share	1193.56
me	719.46	its	1175.01
do	687.78	company	1125.64

How random: I-EN1 vs. I-EN2

- Collect two English Internet corpora 100 MW each
- From two different sets of seeds

More in I-EN2	I-EN1	I-EN2	LL-score	More in I-EN1	I-EN2	I-EN1	LL-score
I	5296.93	6634.12	143.14	tea	17.18	105.07	70.47
June	135.33	380.78	120.60	Christmas	26.66	87.28	34.21
Posted	201.62	455.60	99.64	dog	40.82	101.77	27.17
book	313.75	545.32	62.09	and	21990.16	22902.93	24.01
Definitions	5.16	57.78	51.45	Tea	3.95	29.70	22.37
blog	26.24	105.42	50.74	Speaker	10.33	42.36	21.00
that	8573.25	9555.52	47.98	PST	8.47	37.95	20.34
think	494.23	737.16	47.02	Feb	17.43	54.51	20.21
References	14.92	76.95	45.66	dogs	20.72	59.18	19.46

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`
- WaCky crowd: Marco Baroni, Massimiliano Ciaramita, Silvia Bernardini, Stefan Evert, Bill Fletcher, Adam Kilgarriff, Serge Sharoff. . .

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`
- WaCky crowd: Marco Baroni, Massimiliano Ciaramita, Silvia Bernardini, Stefan Evert, Bill Fletcher, Adam Kilgarriff, Serge Sharoff…
- Web interface(s) and an open source toolkit.

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`
- WaCky crowd: Marco Baroni, Massimiliano Ciaramita, Silvia Bernardini, Stefan Evert, Bill Fletcher, Adam Kilgarriff, Serge Sharoff. . .
- Web interface(s) and an open source toolkit.
- The WaCky philosophy: try to get something concrete out there very soon, so that other will feel motivated to contribute.

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`
- WaCky crowd: Marco Baroni, Massimiliano Ciaramita, Silvia Bernardini, Stefan Evert, Bill Fletcher, Adam Kilgarriff, Serge Sharoff. . .
- Web interface(s) and an open source toolkit.
- The WaCky philosophy: try to get something concrete out there very soon, so that other will feel motivated to contribute.
- 4 100 MW corpora (EN, DE, RU, ZH) available now `http://corpus.leeds.ac.uk/internet.html`

Enter WaCky!

- The **W**eb-**a**s-**C**orpus **k**ool **y**nitiative.
- `http://wacky.sslmit.unibo.it/`
- WaCky crowd: Marco Baroni, Massimiliano Ciaramita, Silvia Bernardini, Stefan Evert, Bill Fletcher, Adam Kilgarriff, Serge Sharoff. . .
- Web interface(s) and an open source toolkit.
- The WaCky philosophy: try to get something concrete out there very soon, so that other will feel motivated to contribute.
- 4 100 MW corpora (EN, DE, RU, ZH) available now `http://corpus.leeds.ac.uk/internet.html`
- 3 1-billion word corpora (English, German, Italian) by spring 2006.