# A Corpus Factory for Indian Languages

Adam Kilgarriff
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Siva Reddy
IIIT Hyderabad, India
gvsreddy@students.iiit.ac.in

Jan Pomikálek
Masaryk Uni., Brno, Cz
xpomikal@fi.muni.cz

Avinesh PVS
IIIT Hyderabad, India
avinesh@students.iiit.ac.in

Lexical Computing Limited

# How is this related?

The morphological systems frequently encounter new words as they are constantly added into the daily language use.

Addition of these new words to a morphological lexicon requires determining their base forms and inflectional paradigms.

To build these morphological lexicons automatically one needs good coverage of the language.

# Problem

Morphological Analyzers and Lexicography
requires large corpora
**but**
many languages ***lack large corpora***

# Outline

- Introduce you to
  - Web Corpora Collection
  - Corpus Factory
- Method of Corpus Building
- Evaluation and Results
- What all can we do with Corpus Factory ??
- Conclusions

# Why Large Corpora ?

- Lexicography benefits from large electronic corpora
  - Assisting the lexicographer.
  - Improving accuracy.
    - Innovations of COBUILD project.
- Statistical NLP
- Theoretical NLP

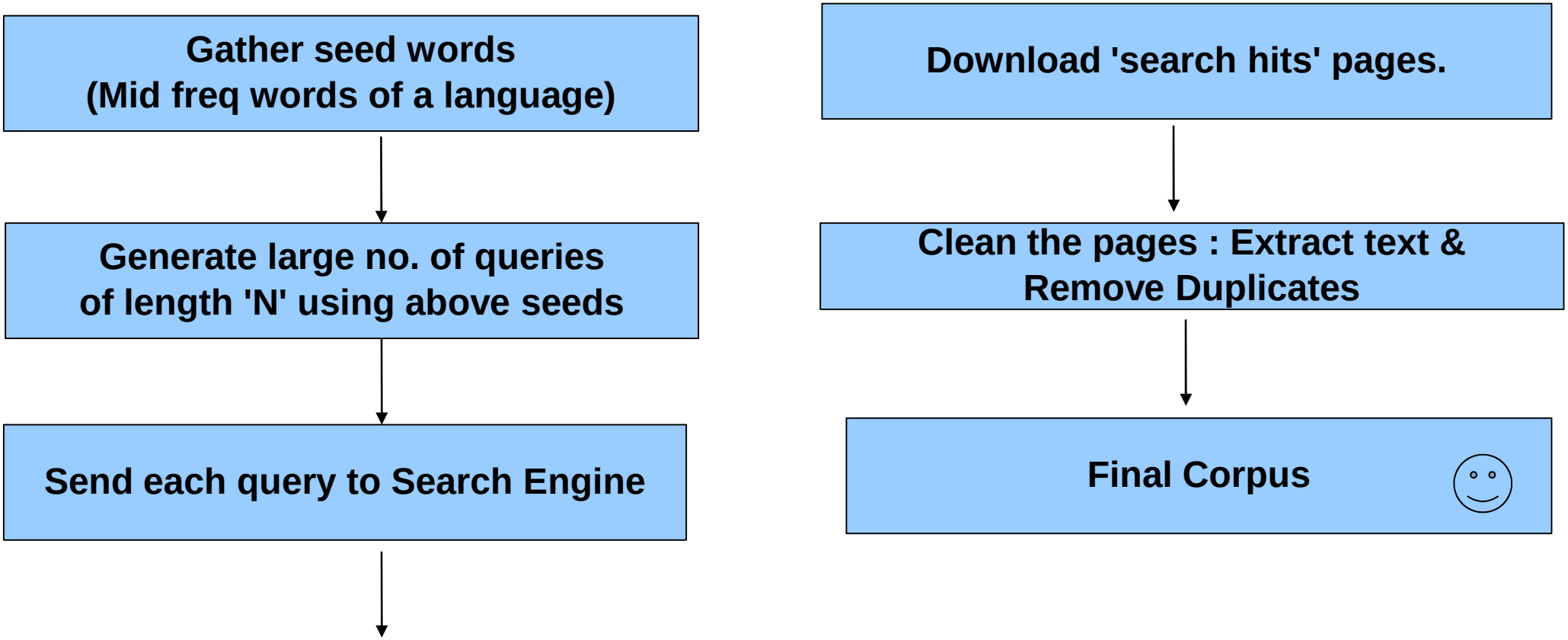
# Web Corpora Collection

- Goal: Building large corpora (100 million words)
- Manual corpus-building
  - slow, labour intensive and expensive
- Solution: Using Web as corpus
- Is the Web a Corpus ?
  - Yes.
  - Question to be asked: Is it a good corpus for the task at hand ?

# Web Corpora Collection: cont..

- Morphology,  Lexicography and NLP
  - Larger the corpus, the better it is.
  - Many Domains
- Why use Web ?
  - Largest source of electronic text.
  - Size of Web.
  - Capture changes in language – New vocabulary.

# Earlier approaches: BooTCat (2004) and Serge Sharroff (2006)



# Corpus Factory

- Minimize human intervention
  - Automate the process of collection
- Goals: Generate any language corpus
  - In a less time
  - With minimal labour, costs
  - Large and Clean Corpus over many domains.
- A factory for building large scale corpus of any language.

# How to achieve Corpus Factory Goals

At each step of
web corpora collection,
we identify the **bottleneck** and solve it

# Step1: Gather Seed Words

- Sharoff used 500 common words drawn from word lists from pre-existing corpora
  - Eg: BNC for English, RNC for Russian
- Bottleneck: No pre-existing large general corpora for many languages.
  - Seed words from many domains required.

# Step1: Gather Seed word

- Wikipedia (Wiki) Corpora : Gather seeds from it
  - Articles from many domains
  - Cheap
  - 265 languages covered : More to come
- Extract text from Wiki.
  - Wikipedia 2 Text
- Tokenise the text.
  - Morphology of the language is important
  - Can use the existing word tokeniser tools.

# Step 1: Gather Seed word

- Build the word frequency lists
  - Sort based on document freq (Descending)
- Top words in the frequency list are the most frequent (Function) words of the lang.
  - Top 500 words (roughly)
  - Helps in identifying connected text.
- We select mid frequent words as seeds
  - $1000^{th}$ to $6000^{th}$ words (roughly)

# Step 2: Query Generation

- Bottleneck: Identifying length of a query
- Shorter length
  - Less number of queries are generated
  - Final corpora size is small.
- Longer Length
  - Data sparsity problems.
  - Less number of hits
- Reasonable length
  - Min hit count of most of the queries is 10.

# Step 2: Query Generation

**Query length, hit counts at 90th percentile and Best Query Length**

| Language | Length= 1 | 2 | 3 | 4 | 5 | Best |
|---|---|---|---|---|---|---|
| **Hindi** | 30,600 | 86 | 1 | - | - | **2** |
| **Telugu** | 668 | 2 | - | - | - | **2** |

# Step 3: Collection

- Around 30,000 queries are generated in the previous step.
- Retrieve top 10 search hits of each query.
  - Yahoo Search API
- Download all pages of search hits.
- Downloaded pages contain unwanted (markup) text.

# Step 4: Cleaning

- Body Text Extraction
  - Bioler plate text like navigation bars, advertisements and other recurring material like legal disclaimers are removed
  - Body Text Extraction algorithm (BTE, Finn et al. 2001)
    - Bioler plate generally is rich in markup
    - Body Text is light in markup.
- Performed on all the downloaded pages to get plain text pages.

# Step 4: Encoding Issues

- Search engines generally normalize different encodings to UTF-8 before indexing.
- The pages downloaded may have different encodings.
- Example: In Hindi, Shusha, Devpooja, Devpriya and UTF-8 are famous.
  - Identify encoding of the page.
  - Normalize it to UTF-8.

# Step 4: Encoding Issues

- Very bad hit on Indian Languages
  - Many encodings <--> font pairs exist
  - Cannot be retreived by UTF-8 query
  - Solution
    - Generate Queries in the native encoding
    - Padma Plugin

# Step 5: Filtering

- Pages may contain unconnected text which is not desirable.
  - Eg: Menu of a hotel, list of names etc.
- Connected text in sentences reliably contains a high proportion of function words (Baroni, 2007)
- We determine the ratio of function words to non function words from wiki corpora.
- Discard the pages if this ratio is not met.

# Step 5: Near Duplicate Detection

- Pages may contain duplicates.
- Duplicates are removed using Broder et al (1997) similarity measure.
  - Two documents are similar if the similarity is greater than a threshold.
  - Similarity is based on the number of overlaps in their n-grams.
- Duplicate pages are removed.

Final Corpus of the
desired language
is obtained.

# Web Corpora Statistics

**Table : Web Corpus Statistics for Different Fonts**

| Language (UTF8) | Unique URLs collected | After filtering | After de-duplication | Web corpus size | |
|---|---|---|---|---|---|
| | | | | MB | Words |
| **Hindi** | 71,613 | 20,051 | 13,321 | 424 MB | 30.6 m |
| **Telugu** | 37,864 | 6,178 | 5,131 | 107 MB | 3.4 m |

| Hindi Font | Unique URLs collected | After filtering | After de-duplication | Web corpus size | |
|---|---|---|---|---|---|
| | | | | MB | Words |
| **Shusha** | 30,322 | 6,052 | 4,985 | 102 MB | 3.2 m |
| **Dev-pooja** | 3,396 | 5,23 | 351 | 10 MB | 0.2 m |

# Evaluation

- Corpus evaluation is a complex matter.
- Good Corpus ??
  - If it supports us in doing what we want to do.
  - Only after using the corpus we get to know.

- The other strategy generally used is by comparison
  - comparing one corpus with another
    i.e. comparing frequency lists of the two corpora

# Evaluation: cont..

- For each of the languages, we have two corpora available:
  - the Web corpus and the Wiki corpus.
- Hypothesis: Wiki corpora are more ‘informational’
  - Informational --> typical written
  - Interactional --> typical spoken
- Lexicography needs more Interactional Corpus.

# Comparing Wiki and Web Corpora

| | **Wiki Corpora (millions of words)** | **Web Corpora (millions of words)** |
|---|---|---|
| **Hindi** | 3.9 | 30.6 |
| **Telugu** | 0.47 | 3.4 |

# Evaluation: cont..

- Goal: Prove Web corpora is more interactional.
- Criticism: Web corpora collected using seeds from wiki corpora (informational) may not be interactional.

# Evaluation: cont..

- First and second person pronouns are strong indicators of interactional language.
  - For English: *I me my mine you your yours we us our*
- Ratio of common 1st and 2nd person pronouns of web and wiki corpora per million corpus are calculated.

# Results

| Telugu | | | |
|---|---|---|---|
| Word | Web | Wiki | Ratio |
| నా | 3736 | 603 | 6.18 |
| నేను | 3390 | 461 | 7.34 |
| నాది | 44 | 17 | 2.59 |
| నన్ను | 585 | 127 | 4.58 |
| మీ | 2092 | 572 | 3.65 |
| మీరు | 1756 | 476 | 3.68 |
| నువ్వు | 281 | 89 | 3.15 |
| మీకు | 730 | 182 | 3.99 |
| నీవు | 80 | 148 | 0.54 |
| నీ | 465 | 263 | 1.76 |
| **Total** | **15755** | **3176** | **4.96** |

# Results

| Hindi | | | |
|---|---|---|---|
| Word | Web | Wiki | Ratio |
| मैं | 2363 | 360 | 6.55 |
| मेरा | 578 | 90 | 6.39 |
| तुम | 827 | 114 | 7.23 |
| आप | 1725 | 664 | 2.59 |
| आपका | 192 | 54 | 3.50 |
| मैंने | 709 | 65 | 10.76 |
| मुझे | 1404 | 122 | 11.50 |
| तू | 185 | 50 | 3.65 |
| तुम | 827 | 114 | 7.23 |
| तूने | 23 | 12 | 1.85 |
| **Total** | **8833** | **1645** | **5.36** |

Results prove that
Web Corpora is more interactional.

# What all can we do with Corpus Factory

- Platform to create any language resource.
- Build Dictionaries with much less efforts.
  - Corpus Factory + Sketch Engine
  - Sketch Engine
    - Most powerful and mostly used concordance Engine
    - Corpus Analysis tools
- University/Individual Licences available.

# What all can we do with Corpus Factory

- Impact on current state of art of many languages.
- Statistical Methods
  - Corpus is the first step to start with and you have it.
  - Any corpus based method is only as good as the corpus it uses
- Future Goal: Provide resources to as many languages as possible.

# Hindi in Sketch Engine

| | |
|---|---|
| 3742.html | की जमीन रामगढ़ और बिल्ला गांव के बीच जंगलनुमा |
| 14710.html | सदस्य हाथ में काला बिल्ला लगाये हुए थे । इसके |
| 21899.html | कार्लिश्टफ्ट के बिल्ला सुपर स्टोर में |
| 56617.html | पता नहीं कौन सा बिल्ला चट कर गया था ।  |
| 56617.html | चुपचाप खड़ा रहा । बिल्ला कुछ देर जवाब का |
| 56617.html | यह मुगालता था कि बिल्ला उसका रास्ता छोड़ |
| 56617.html | और कुर्ता पहने एक बिल्ला खड़ा था । वह इस अंदाज |
| 56617.html | लेकिन यह तो सचमुच बिल्ला था । उसने अपने कुर्ते |
| 56617.html | गुर्राहट सुन कर बिल्ला पीछे मुड़ा । उसने |
| 56617.html | और इससे पहले कि बिल्ला कुछ समझ पाये , उसके |
| 56617.html | वाले इस हमले से बिल्ला जरा भी परेशान नहीं |
| 56617.html | उठाये गये थे , अब एक बिल्ला बैठा था । उसने इस |
| 56617.html | हौलनाक कारनामे के बाद बिल्ला एक कमजोर बूढ़े आदमी |
| 56617.html | भलमनसाहत थी , जिसे देख कर बिल्ला खीज उठा । उसने मुंडी |

# Telugu in Sketch Engine

| | | | |
|---|---|---|---|
| 15231.txt | బాల్ బాయ్స్ . అక్కడ | బంతి | అందించే వ్యక్తిగా |
| 9670.txt | నాకివ్వని వాడి | బంతి | పగిలిపోయి ఎగరనప్పట్లా |
| 25124.txt | మ్యాచ్ లో చివరి | బంతి | కి ఫోరు కొట్టి |
| 9149.txt | ఇప్పటికీ ముద్ద | బంతి | పూవు లాంటి తెలుగు |
| 9335.txt | ఇనుము-రాయి తో చేసిన | బంతి | తన గరిమ బలాన్ని |
| 9335.txt | క్షేత్రంలో జరిగితే అక్కడ | బంతి | ఫోటాను అన్న మాట |
| 12463.txt | అభివర్ణించాడు . | బంతి | గతి మారుతుంది . |
| 19783.txt | తీయాలనుకోండి . ఒక | బంతి | ఎగురుతున్నట్టు |
| 19783.txt | దానికి ముందుగా | బంతి | వేరు వేరు స్థాయిలలో |
| 19783.txt | గీసుకోవాలి . అంటే | బంతి | ఎగురుతున్నప్పుడు |
| 19783.txt | అయిపోతుందనుకున్నారా ? .  ఆ | బంతి | ఒకవేళ నేలకు తగిలి |
| 19783.txt | నేలకు తాకినప్పుడు ఆ | బంతి | ఏ మేరకు కుచించుకుపోతుంది |
| 35356.txt | ఎక్కువగా పూస్తాయి . | బంతి | , చేమంతి , నంది వర్ధనం |
| 7076.txt | పద్యాలు .  ముద్ద | బంతి | పూవులో&#8230 ; ( మూగ |
| 15911.txt | సాధారణం అయిపోయింది . ఆ | బంతి | చుట్టుపక్కల ఇళ్ళలో |

# Future Work

- Prepare corpora for Indian languages, Korean, Tibetan and all the official languages of the European Union.
- Encoding problems
  - Bad hit on Indian languages.
  - Extensive survey on Thai, Tibetan and Vietnamese.
- Estimate the Size of Web
- Bing Vs Google Vs Yahoo

# Conclusions

- Corpus Factory presents
  - A method for developing large general-language corpora which can be applied to many languages
  - Large corpora for five languages currently.
  - Many to come
    - a great deal to offer **Asian lexicography**

# Thank you

Lexical Computing Ltd,
http://sketchengine.co.uk

inquiries@sketchengine.co.uk