An $n$-gram cache for large-scale parallel extraction of multiword relevant expressions with LocalMaxs

Carlos Gonçalves$^{1,2}$; Joaquim F. Silva$^2$; José C. Cunha$^2$

$^1$ Instituto Superior de Engenharia de Lisboa, Portugal
$^2$ NOVA LINCS, Faculdade Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Presented by Carlos Gonçalves (cgoncalves@deetc.isel.pt)

Acknowledgments: FCT MCTES, NOVA LINCS UID/CEC/04516/2013; Lunacloud (http://www.lunacloud.com); ISEL “AzoresCloud”
Agenda

- Objective
- LocalMaxs
- Distributed Architecture
- $n$-grams Statistical Distribution
- $n$-gram Cache System
- Experimental Results
- Conclusions and Further Work
Statistical Extraction of Topics

Time

Scale

Cloud / Cluster
LocalMaxs: Statistical Extraction

Relevance \equiv \text{Strong internal co-occurrence (glue) of words}

\[ SCP_f(w_1 \ldots w_n) = \frac{p(w_1 \ldots w_n)^2}{\frac{1}{n-1} \sum_{i=1}^{n-1} p(w_1 \ldots w_i) \ast p(w_{i+1} \ldots w_n)} \]

Detecting local maxima of the glue

J. F. da Silva and G. P. Lopes. “A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units”
LocalMaxs: Relevance Criterion

Given a corpus, LocalMaxs algorithms extracts a set of multiword relevant expressions

A) Bigrams case: length(W) = 2

Comparing to adjacent enclosing \((n+1)\)-grams \(\rightarrow 3\)-grams

\[\Omega_{n+1}(W)\]

\[\begin{align*}
\text{the} & \quad \text{energy saving} & \quad \text{has} \\
\text{in} & \quad \text{} & \quad \text{trust}
\end{align*}\]

\[\forall \text{ glue } y \in \Omega_{n+1}(W)
\]

\[SCP(W) > y \rightarrow W \text{ is Relevant!}\]
LocalMaxs: Relevance Criterion

B) Higher $n$-grams: $\text{length}(W) > 2$

Comparing to the adjacent enclosing $(n+1)$-grams

$$\Omega_{n+1}(W)$$

The diagram shows a string $W$ which is longer than 2 characters. The context includes phrases like "energy saving in the public sector" and "has cover."
LocalMaxs: Relevance Criterion

B) Higher $n$-grams: $\text{length}(W) > 2$

Comparing to the adjacent enclosed $(n-1)$-grams $\Omega_{n-1}(W)$

- energy saving in the public sector
- energy saving in the public
LocalMaxs: Relevance Criterion

B) Higher \( n \)-grams: \( \text{length}(W) > 2 \)

Comparing to the adjacent enclosed \((n-1)\)-grams

\[ \Omega_{n-1}(W) \]

energy saving in the public sector

saving in the public sector
LocalMaxs: Relevance Criterion

B) Higher $n$-grams: $\text{length}(W) > 2$

Comparing to the adjacent $n$-grams

for all glue $x \in \Omega_{n-1}(W)$, for all glue $y \in \Omega_{n+1}(W)$

$$SCP(W) > \frac{x + y}{2}$$

$\Rightarrow W$ is Relevant!
LocalMaxs

1) Count all \textit{n-gram} frequencies in the \textit{corpus}
2) Calculate all distinct \textit{n-gram} glues (cohesion)
3) Find Relevant Expressions: for \textit{n-gram} sizes from 2 to N, select the local stronger average glues

**Approaches**

<table>
<thead>
<tr>
<th>Sequential</th>
<th>Parallel &amp; Distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very time-consuming !!!</td>
<td>Parallel: To reduce Time</td>
</tr>
<tr>
<td>Huge memory-demanding !!!</td>
<td>Distributed: To fit in Memory</td>
</tr>
</tbody>
</table>
Distributed Architecture

- Generic architecture capable of executing algorithms based on statistical $n$-gram models

- Capable of being executed in cluster or cloud environments
Distributed Architecture

LocalMaxs

functions
Distributed Architecture

In-memory

Key Value Store

Interconnection
Distributed Architecture

n-gram

Cache system
Distributed Architecture

- Phase 1 counts the $n$-gram occurrences
  - Distributed hash table with the $n$-gram data
- Phase 2 calculate the cohesion
- Phase 3 identifies the $n$-grams that can be considered Relevant Expressions
Distributed Architecture

• Phase 1 counts the $n$-gram occurrences
  - Distributed hash table with the $n$-gram data

• Phase 2 calculate the cohesion

• Phase 3 identifies the $n$-grams that can be considered Relevant Expressions
Distributed Architecture

- Phase 1 counts the $n$-gram occurrences
  - Distributed hash table with the $n$-gram data
- Phase 2 calculate the cohesion
- Phase 3 identifies the $n$-grams that can be considered Relevant Expressions
**n-gram Cache System**

- An *n*-gram cache system, to reduce the remote data communication
- Analytical model to understand cache miss ratio and miss penalty
- Cooperative warm-up strategy
- Finite size or *infinite*, depending on algorithm requirements or system resources

---

**Diagram:**
- **KVS Server**
- **Controller** (multi thread)
- **n-gram Cache System**: $C_1, C_2, \ldots, C_n$

**Text:**
- Enough to contain all distinct *n*-grams
\( n \)-grams Statistical Distribution

- \( n \)-gram repetition depends on:
  - Corpus size
  - Language
  - \( n \)-gram size

- Empirical and Theoretical approaches

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
  & 1-gram & 2-gram & 3-gram & 4-gram & 5-gram & 6-gram \\
\hline
\( D_i \) & 1.95\times10^8  & 7.08\times10^8  & 3.54\times10^9 & 9.80\times10^9 & 5.06\times10^{10} & 3.92\times10^{11} \\
 & (0.2 \text{ Gw}) & (0.7 \text{ Gw}) & (3.5 \text{ Gw}) & (9.8 \text{ Gw}) & (50 \text{ Gw}) & (0.4 \text{ T w}) \\
\hline
\( |C| \) & 9.22\times10^{11} & 1.05\times10^{12} & 1.29\times10^{12} & 1.43\times10^{12} & 6.18\times10^{12} & 2.39\times10^{13} \\
 & (0.9 \text{ T w}) & (1 \text{ T w}) & (1.2 \text{ T w}) & (1.4 \text{ T w}) & (6.2 \text{ T w}) & (24 \text{ T w}) \\
\hline
\end{tabular}
\end{center}

English plateaux

Plateaux for the English language
Experimental Results

- Multiple runs in public cloud (Lunacloud): virtual machines: 4 CPU@1.5 GHz and one local partition of 10 Gbyte

- Different number of machines (1, 9, 16, 24, 32, 40 and 48) with RAM ranging from 16 to 90 Gbyte, and different corpus sizes (25, 227, 466 and 682 million words)

- Evaluate:
  - \( n \)-gram cache evaluation – Real execution results vs model estimates;
  - LocalMaxs phase 2 real execution time and cache behavior;
  - LocalMaxs total execution time vs phase 2 execution time
Experimental Results, 16 up to 48 virtual machines
Corpus up to 682 Mw, 2-grams & 3-grams

- Phase two execution time is dominated by the communication due to the \( n \)-gram misses in the observed range of *corpora* size and number of machines
Experimental Results, 16 up to 48 virtual machines
Corpus up to 682 Mw, 2-grams & 3-grams

- Cache miss ratio ≈ 30%

- As corpus sizes increases the miss ratio decreases due to the repetition of n-grams
Experimental Results
Extraction of relevant 2-grams and 3-grams

Fixed-"corpus size"

\[ T \propto \frac{1}{K} \]

\[ S_{p_{K_1 \rightarrow K_2}} = \frac{T_{par}(K_1)}{T_{par}(K_2)} \]
Experimental Results
Extraction of relevant 2-grams and 3-grams

Fixed-time Sizeup

Scale

$$S_{z p K_1 \rightarrow K_2} = \frac{N_{par}(T, K_2)}{N_{par}(T, K_1)}$$
Conclusions and Further Work

• The approach is scalable to larger corpora sizes and higher size $n$-grams by simply increasing the number of machines.

• An $n$-gram cache significantly reduced the remote data communication.

• For each corpus size the number of distinct $n$-grams imposes a limit to the minimum remote communication overhead.
Thank you for your attention

Carlos Gonçalves (cgoncalves@deetc.isel.pt)