Induced acyclic subgraphs with optimized endpoints

Learning strategies

Moussa Abdenbi, Alexandre Blondin Massé & Alain Goupil

Moussa Abdenbi
Université du Québec à Montréal
abdenbi.moussa@courrier.uqam.ca
About me

- My name is Moussa Abdenbi.
- I am a Ph.D student at Université du Québec à Montréal.
- I work on graph theory, specifically on finding induced and acyclic subgraphs of a directed graph.
- I am also working on applications of my research in computational linguistics.
Introduction
Motivation

- Learning a new language or acquiring specialized vocabulary
  - Direct learning: seeing, hearing, smelling, tasting, touching, interacting, ...
  - Learning by definition: reading a definition or being explained something

- Learning a word by definition is easier than learning word directly

Question

How to maximize the learning by definition approach?
Context

- A *strategy* is an ordered sequence of words
  - Carefully choose these words
- Some psycholinguistic criteria
  - The age the word is acquired
  - Rate of occurrence in a given corpus
- In this work, we provide a way to select strategy by using graph theory
  - We focus on strategies for learning by definition approach
Preliminaries
Recall some definitions of graph theory:

- **Directed graph** (*digraph*) $D = (V, A)$
  - $V$ is a set of vertices
  - $A$ is a set of arcs (ordered pairs of vertices)

- **Subgraph** $I = (V_I, A_I)$ of $D = (V, A)$
  - $V_I \subseteq V$
  - $A_I \subseteq \{(u, v) \in A \mid u, v \in V_I\}$
  - It is **induced** if $A_I = A \cap V_I \times V_I$, denoted by $I = D[V_I]$.

- **Path** between $u_0 \neq u_k$ is a sequence $p = (u_0, u_1, \ldots, u_k)$ such that $0 \leq i \leq k - 1$, $(u_i, u_{i+1}) \in A$ or $(u_{i+1}, u_i) \in A$
  - $p$ is called **directed**, if $0 \leq i \leq k - 1$, $(u_i, u_{i+1}) \in A$
Vocabulary (2/2)

- **Circuit** is a directed path $p = (u_0, u_1, \ldots, u_k)$ where $(u_k, u_0) \in A$

- **Connected graph** $D = (V, A)$ if there is a path between any two vertices

- **Acyclic** digraph $D$ if it has no circuit

- **Degree** of $u \in V$
  - Out-degree $\deg_D^+(u) = |\{v \in V \mid (u, v) \in A\}|$
  - In-degree $\deg_D^-(u) = |\{v \in V \mid (v, u) \in A\}|$

- **Source or sink** $u \in V_I$,
  - if $\deg_D^-(u) = 0$ then $u$ is a source
  - if $\deg_D^+(u) = 0$ then $u$ is a sink
Examples

(11, 5, 1, 7) is a directed path and (0, 5, 1, 0) is a circuit. \( I \) and \( J \) are subgraphs.

- \( I \) is not induced:
  - Arc (11, 5) is not in \( I \), when 11 \( \in \) \( I \) and 5 \( \in \) \( I \).

- \( \text{deg}_D(11) = \text{deg}_D^+(11) + \text{deg}_D^-(11) = 3 + 1 \)
- \( \text{deg}_I(11) = \text{deg}_I^+(11) + \text{deg}_I^-(11) = 2 + 0 \)

- \( J \) is induced and acyclic:
  - Vertex 11 is a source and vertex is 7 is a sink in \( J \).
Lexicon
Digraph dictionary

- Dictionary as a directed graph
  - Each word on the dictionary is a vertex in the digraph
  - Arc \((w_1, w_2)\) if and only if \(w_1\) appears in the definition of \(w_2\)
Strategies

- Basically a strategy is a subgraph:
  1. There is no cyclic words definition
     → The subgraph is acyclic
  2. Select words that appear in a large number of definitions
     → Maximize the difference between sinks and sources
  3. Pick all arcs between chosen words
     → The subgraph is induced
  4. Focus learning on a specialized vocabulary
     → The subgraph must contain a fixed set of vertices

- Induced acyclic subgraphs with optimized endpoints
Complexity of the problem
Mathematical formulation

- $S_D$ the set of all subgraphs of $D = (V, A)$
- $\Delta(I) = p_I - s_I$
  - $s_I$ is the number of sources
  - $p_I$ its number of sinks

**Optimization criterion**

Maximize the function $\Delta$ for induced and acyclic subgraph of size $1 \leq i \leq |D|$
Problem formulation

**Decision problem**

Given a digraph $D = (V, A)$, a set of vertices $M \subseteq V$ and two integers $i$ and $\delta$, does there exist an induced and acyclic subgraph of $D$ of size $i$ containing $M$, such that $\Delta(I) = \delta$?

→ NP-complete

**Optimization problem**

Given a digraph $D = (V, A)$ and $M \subseteq V$, what is the maximal value $\Delta(I)$ that can be realized by an induced and acyclic subgraph $I$ of $D$ of size $i$ and containing $M$, for $i \in \{|M|, |M| + 1, \ldots, |D|\}$?

→ NP-hard
Algorithms
Greedy algorithm

- Add the most *interesting* vertices
  - Starting with $I = D[M]$ until $|I| = i$
- The variation they bring to $\Delta(I)$
  - The greater the value a vertex brings, the greater its interest

Tabu algorithm

- Increase $\Delta(I)$ by browsing *neighborhoods* of $I$
- A *neighbor* of $I$ is an induced and acyclic subgraph $I'$, such that,
  - $M \subset V_{I'}$
  - $|I| = |I'| = i$
  - $|V_I \cap V_{I'}| = |V_I| - 2 = |V_{I'}| - 2$
Experimentation
Dictionaries

- The Wordsmyth Illustrated Learner’s Dictionary (WILD)
  → 4244 vertices and 59478 arcs
- The Wordsmyth Learner’s Dictionary-Thesaurus (WLDT)
  → 6036 vertices and 29735 arcs
- The Wordsmyth Children’s Dictionary-Thesaurus (WCDT)
  → 20128 vertices and 107079 arcs

Costs

- Learning by definition, if we know all the words occurring in a definition, then cost is 0
- 1 otherwise
Psycholinguistic strategies

- Brysbaert-AOA: words ordered according to their age of acquisition
- Brysbaert-Concreteness: words ordered from most concrete to most abstract
- Brysbaert-Frequency: words ordered by their rate of occurrences
- Childes-AOA: words from Child Language Data Exchange System project, ordered with respect to their age of acquisition
- Frequency-NGSL: word lists designed and ordered to help students learning English
Graph theory based strategy

- Algorithms with $M = \emptyset$
  - Measure correlation between cost and optimization criterion $\Delta$
  - Psycholinguistic strategies are designed to learn an entire language

- Induced acyclic subgraphs with optimized endpoints

- Vertices of subgraphs considered as a strategy

- Ordered according to their out-degree, from highest to lowest
Comparison criteria

- **cost**: total number of words learned directly
- **efficiency**: ratio of number of words learned over number of words learned directly

Subgraphs computation

- Large size dictionary digraphs
  - Metaheuristics.
  - Greedy algorithm solution as input to tabu search
## Results

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Subgraph strategy</th>
<th>Childes</th>
<th>Freq.</th>
<th>Brysbaert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AOA</td>
<td>NGSL</td>
<td>AOA</td>
</tr>
<tr>
<td>WILD</td>
<td>size</td>
<td>50</td>
<td>1981</td>
<td>1169</td>
</tr>
<tr>
<td></td>
<td>cost</td>
<td>3013</td>
<td>3174</td>
<td>3079</td>
</tr>
<tr>
<td></td>
<td>effic.</td>
<td>1.40</td>
<td>1.33</td>
<td>1.37</td>
</tr>
<tr>
<td>WLDT</td>
<td>size</td>
<td>200</td>
<td>1580</td>
<td>697</td>
</tr>
<tr>
<td></td>
<td>cost</td>
<td>917</td>
<td>1803</td>
<td>1088</td>
</tr>
<tr>
<td></td>
<td>effic.</td>
<td>6.58</td>
<td>3.34</td>
<td>5.54</td>
</tr>
<tr>
<td>WCDT</td>
<td>size</td>
<td>300</td>
<td>3316</td>
<td>1122</td>
</tr>
<tr>
<td></td>
<td>cost</td>
<td>2431</td>
<td>4366</td>
<td>2777</td>
</tr>
<tr>
<td></td>
<td>effic.</td>
<td>12.23</td>
<td>6.81</td>
<td>10.70</td>
</tr>
</tbody>
</table>

- Subgraphs strategies are better
- Subgraphs strategies sizes are smaller than psycholinguistic strategies sizes
Conclusion
- New problem, difficult to compare results
- Even with approximate solutions, learning strategies are better
- Further investigate linguistic applications
- Try other digital dictionaries
Thank you!