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

IARIA

UQÀM

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
  - $\rightarrow$ Direct learning: seeing, hearing, smelling, tasting, touching, interacting, ...
  - $\rightarrow$ 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
  - $\rightarrow$ Carefully choose these words
- Some psycholinguistic criteria
  - $\rightarrow$ The age the word is acquired
  - $\rightarrow$ Rate of occurrence in a given corpus
- In this work, we provide a way to select strategy by using graph theory
  - $\rightarrow$ We focus on strategies for learning by definition approach

Preliminaries

Vocabulary (1/2)

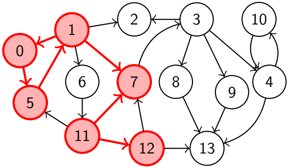
Recall some definitions of graph theory :

- **Directed graph** (*digraph*) $D = (V, A)$
  - $\rightarrow$ V is a set of vertices
  - $\rightarrow$ A is a set of arcs (ordered pairs of vertices)
- **Subgraph** $I = (V_I, A_I)$ of $D = (V, A)$
  - $\rightarrow$ $V_I \subseteq V$
  - $\rightarrow$ $A_I \subseteq \{(u, v) \in A \mid u, v \in V_I\}$
  - $\rightarrow$ 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 $(u_{i+1}, u_i) \in A$
  - $\rightarrow$ p is called **directed**, if $0 \leq i \leq k - 1$, $(u_i, u_{i+1}) \in A$

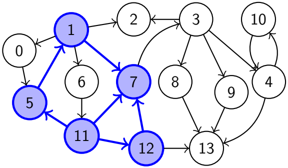
Vocabulary (2/2)

- **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



I



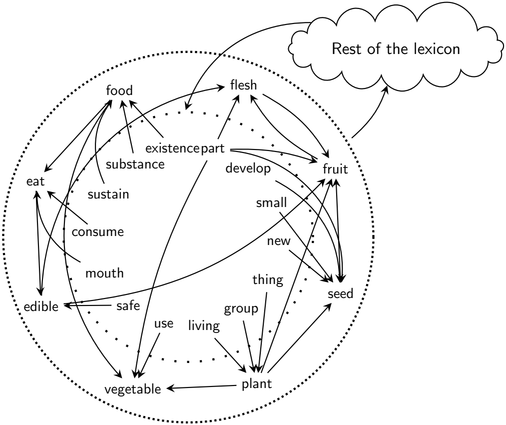
J

- $(11, 5, 1, 7)$ is a directed path and $(0, 5, 1, 0)$ is a circuit. I and J are subgraphs
- I is not induced :
  - $\rightarrow$ Arc $(11, 5)$ is not in I, when $11 \in I$ and $5 \in I$.
- $deg_D(11) = deg_D^+(11) + deg_D^-(11) = 3 + 1$
- $deg_I(11) = deg_I^+(11) + deg_I^-(11) = 2 + 0$
- J is induced and acyclic :
  - $\rightarrow$ Vertex 11 is a source and vertex is 7 is a sink in J.

Lexicon

Digraph dictionary

- Dictionary as a directed graph
  - $\rightarrow$ Each word on the dictionary is a vertex in the digraph
  - $\rightarrow$ 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

     $\rightarrow$ The subgraph is acyclic
  2. Select words that appear in a large number of definitions

     $\rightarrow$ Maximize the difference between sinks and sources
  3. Pick all arcs between choosen words

     $\rightarrow$ The subgraph is induced
  4. Focus learning on a specialized vocabulary

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

Optimization criterion

Maximize the function Δ 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 $M \subset V$ and two integers i and δ, does there exist an induced and acyclic subgraph of D of size i containing M, such that $\Delta(I) = \delta$?

$\rightarrow$ NP-complete

Optimization problem

Given a digraph $D = (V, A)$ and $M \subset V$, 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 $i \in \{|M|, |M| + 1, \ldots, |D|\}$?

$\rightarrow$ NP-hard

Algorithms

Greedy algorithm

- Add the most *interesting* vertices
  - $\rightarrow$ Starting with $I = D[M]$ until $|I| = i$
- The variation they bring to $\Delta(I)$
  - $\rightarrow$ 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,
  - $\rightarrow$ $M \subset V_{I'}$
  - $\rightarrow$ $|I| = |I'| = i$
  - $\rightarrow$ $|V_I \cap V_{I'}| = |V_I| - 2 = |V_{I'}| - 2$

Experimentation

Dictionaries

- The Wordsmyth Illustrated Learner's Dictionary (WILD)
  - $\rightarrow$ 4244 vertices and 59478 arcs
- The Wordsmyth Learner's Dictionary-Thesaurus (WLDT)
  - $\rightarrow$ 6036 vertices and 29735 arcs
- The Wordsmyth Children's Dictionary-Thesaurus (WCDT)
  - $\rightarrow$ 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 $M = \emptyset$
  - $\rightarrow$ Measure correlation between cost and optimization criterion Δ
  - $\rightarrow$ 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

  $\rightarrow$ Metaheuristics.

  $\rightarrow$ Greedy algorithm solution as input to tabu search

Results

Dictionary		Subgraph strategy	Childes AOA	Freq. NGSL	Brysbaert AOA	Brysbaert Concret.	Brysbaert Freq.
WILD	size	50	1981	1169	1369	2417	1188
	cost	**3013**	3174	3079	3079	3207	3059
	effic.	**1.40**	1.33	1.37	1.37	1.32	1.38
WLDT	size	200	1580	697	1277	2366	957
	cost	**917**	1803	1088	1530	2485	1296
	effic.	**6.58**	3.34	5.54	3.94	2.42	4.65
WCDT	size	300	3316	1122	2879	5974	1995
	cost	**2431**	4366	2777	4016	6616	3315
	effic.	**12.23**	6.81	10.70	7.40	4.49	8.96

- 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!