Dictionaries and Lexical Graphs

Gary Coen

Boeing Phantom Works, M&CT
PO Box 3707 MC 7L-43
Seattle WA 98124-2207
Gary.A.Coen@boeing.com

Abstract: This paper introduces lexical graphs, which model dictionaries in graph-theoretic terms. After briefly reviewing some basic tenets of graph theory, lexical graphs are formally described along with techniques for (*i*) inducing a topology on a dictionary, (*ii*) measuring lexical stability in a dictionary, and (*iii*) distinguishing between core and peripheral information concepts encoded in a dictionary. These are useful results: the sorted topology of a dictionary provides a learning sequence for the subject matter compiled in it, and lexical stability identifies the contribution of a particular lexical entry to the structural integrity of the dictionary as a whole. Exploiting these and other properties, it is possible to identify a lexical stability value for a single dictionary entry, a subset of interdependent dictionary elements, and the dictionary as a whole. Hence, lexical graphs can play a significant role in the analysis of dictionary information structure, thus benefiting natural language engineering technologies of various kinds.

1 Graphs

A graph G is an ordered pair (V, E), where V is a finite set of vertices and E is a binary relation on V composing a set of edges. An edge is a pair (u, v) with $u, v \in V$. If edge $e = (u, v)$ is in graph G, then u and v are said to be the end vertices of e.

An edge is directed if its end vertices are an ordered pair. Suppose that e is an out-going edge of u and an incoming edge of v. A specialized edge like e is an arc, and its specialized end vertices are nodes. A directed graph is a graph in which all edges are arcs. If arcs $e=(u, v)$ and $e'=(u, w)$ exist in a directed graph, then e and e' are incident to node u. The degree of node u, $d(u)$, is the total number of arcs incident to u. Moreover, the incoming degree of u, $d^{-}(u)$, is the total number of incoming arcs incident to u, and the outgoing degree of u, $d^{+}(u)$, is the total number of outgoing arcs incident to u.

In a directed graph $G=(V, E)$, a walk is a sequence between end nodes of one or more arcs, and a path is a walk in which all nodes are distinct. Node v of G is reachable from node u if $v = u$ or G contains a path from u to v. G is connected if there exists a path between every pair of nodes in G; otherwise, G is disconnected. A subgraph S of G has

all its arcs and nodes in G. A directed subgraph S of G rooted at v is a directed graph $S=(V', E')$ where $v \in V$, $V' \subseteq V$, $E' \subseteq E$, and V' is the set of nodes from which v is reachable. Finally, a cycle is a walk in which the first and last nodes are identical while all others are distinct, and a directed acyclic graph is a directed graph without cycles.[1]

2 Lexical Graphs

Let D be a dictionary, a finite set of strings arranged as ordered pairs (t, d) such that each d is said to define its t. A lexical graph is a DAG $L=(D, E)$ where each pair in D is a node in L, and E is a binary relation on D. Since the end nodes u, v of each arc in E are also ordered pairs, a more precise notation for an arc in a lexical graph is needed. For convenience, let each arc in E be annotated as $e=((t, d), (t', d'))$ then whenever $e \in E$ of L, the definition d can be said to depend on the term t'.[2]

2.1 Lexical Dependency

Applying the foregoing description of lexical graph L, suppose lexical dependency obtains between nodes (t, d), (t', d') whenever t' is a substring of d in the arcs of L.[3] Then the lexical dependency relation thus defined on L imposes a topology on its dictionary. Each (t, d) pair in D of L identifies the information structure of a term under definition, and whenever an element in D lexically depends on another to ground its meaning, a corresponding arc exists in E. Hence, lexical dependency intrinsically orders D: whenever $e=((t, d), (t', d')) \in E$ of L, it is clear that (t', d') precedes (t, d) in the topology of D. This observation permits the following induction: knowledge of the meaning of an element in a dictionary topology cannot be guaranteed without prior knowledge of the meaning of the preceding elements in that topology.

2.2 Lexical Stability

Intuitively, the stability property of a particular node in a lexical graph is a discrete measure of its contribution to the structural integrity of that graph.[4] In lexical graph L, the stability of v, $v \in D$, expresses the potential impact of any modification to v on the configuration of L. To illustrate, consider (1) from Figure 1, a simple lexical graph of three nodes x, y, z:

---

[1] Directed acyclic graphs are often called DAGs. For a thorough introduction to graph theory, try [Wa72].

[2] Thus dependency information in lexical graphs identifies the same information as that captured by dependency graphs, for instance, when used with attribute grammars to validate, optimize, and translate computer programs. The literature in this area is extensive; [Kn68] establishes the paradigm.

[3] The substring relation suffices for expository purposes. Of course, lexical dependency can be formulated in terms of other relations as well.

[4] Robert Martin (personal communication) introduced me to the formal study of stable interdependencies among information units within the context of build dependencies in the programming language C++. (Martin acknowledges Bertrand Meyer as the originator of the concept.) For details of this problem and its treatment, see Martin's *Engineering Notebook* columns of *The C++ Report* for the year 1999. The theory of lexical graphs stems in part from this inspiration, although any errors are mine alone.

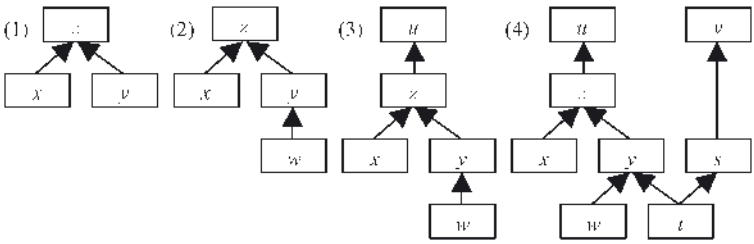

Figure 1. Four Simple Lexical Graphs

Nodes x and y lexically depend on z, and z is lexically independent. Clearly, the meaning of x and y can change without affecting the meaning of z. Moreover, no change in x will affect the meaning of y, and vice-versa. However, a change to z may reformulate the meaning of both x and y, transforming the structure and content of L to an extent measured by the stability of z. Alternatively, a potential side effect of change to nodes x and y, which follow z in the dictionary topology, is that they may no longer follow z afterward, depending on the nature of the modification to dictionary elements x and y.

2.2.1 Measuring Lexical Stability

Interdependencies between nodes in a lexical graph determine the global information structure of the graph as well as the local information structure of each node. Recall that $d(v)$, the degree of node v, is the total number of arcs incident to v. Likewise, $d^-(v)$, the incoming degree of v, summarizes the incoming arcs of v. The lexical stability of node v, $S(v)$, is then the quotient of dividend $d^-(v)$ and divisor $d(v)$:

Definition 1: Let v, $v \in D$, be a node in lexical graph $L = (D, E)$. Then $S(v) = \dfrac{d^-(v)}{d(v)}$.

Application of this measurement to lexical graph (1) of Figure 1 indicates $S(z) = 1$, the maximal value for lexical stability, while $S(x)$ and $S(y)$ both yield 0, the minimal value. As lexical stability approaches the maximum, change to the information concept defined in a dictionary node has more pervasive effects on the structural configuration of the lexical graph. As a practical matter, a high stability value local to some node v implies difficulty of change for the dictionary entry encoded at v, since it and each of the entries encoded in the subgraph rooted at v must be reviewed for correctness subsequent to the change. Conversely, a reduced burden of validation accompanies modification of a node with lower lexical stability. Hence, low stability implies ease of change.

Consistent with this framework of evaluation, the lexical stability metric assigns maximum and minimum stability values to the root and leaf nodes, respectively, of a lexical graph. This is an intuitive result. By fiat, orphans (*i.e.*, isolated lexical graph nodes where $d(v)=0$) possess minimum stability.

Next consider lexical graph (2) from Figure 1, which extends (1) such that node w depends on y. Again, the calculation of $S(v)$ identifies minimal stability for the leaf

nodes of the new configuration. Node w depends on y and y depends on z, and the calculation identifies a higher lexical stability value for z than y: S(z)=1 and S(y)=0.5. Moreover, each lexical dependency added to node y increments S(y) appropriately, although S(y) will never achieve the maximum value. Hence, S(z) will always exceed S(y) in lexical stability, another intuitively appropriate result. Now consider lexical graph (3), which extends (2) such that z, the former root node, has been made to depend on u, a new root node. In this configuration, S(z) is devalued to 0.66 and S(u)=1. Nodes y, w, and x are the dictionary entries easiest to change, and they retain their previous lexical stability values.

Finally consider lexical graph (4), which extends (3) by adding a dependent node t to y and making t dependent on an additional node s. The new s node, in turn, depends on a third additional node v. As in lexical graphs (1)-(3), the two root nodes of (4) have maximal stability. Nodes x and w remain leaf nodes, and thus retain minimum lexical stability. S(y) has increased to 0.66, reflecting the additional lexical dependent t. Node z presents the same configuration as y, hence S(z) 0.66. Node s has identical incoming and outgoing degrees; hence, its lexical stability is 0.5. Figure 2 summarizes:

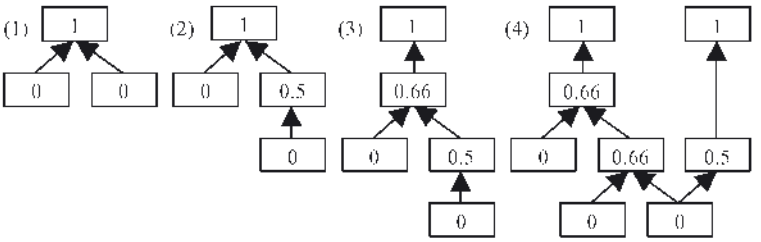

Figure 2. Lexical Stability in Four Simple Lexical Graphs

2.2.2 Lexical Stability and Information Structure

Measurement of lexical stability is local to a node in a lexical graph. Should modification of a node change its relationship to another node, the topology of the dictionary may change. Not every change has structural consequences. A change in a leaf node or an orphan may be trivial, as the minimum lexical stability value of such nodes attests. On the other hand, some changes may have intricate consequences throughout the dictionary. Modification of a node with high incoming degree positioned near a root, for instance, may have a significant impact on the information structure of a dictionary.

Formally, modification of lexical graph L is accomplished by creating, updating, or deleting one or more elements in D of L. Suppose two nodes v (t, d) and w (t', d') exist, where v, $w \in D$ and v lexically depends on w. Recall that whenever a node lexically depends on another to ground its meaning, a corresponding arc $e=((t, d), (t', d'))$ appears in E of L. Should v change such that it no longer depends on w, the result is that $e \notin E$ of L. Should v change such that it retains its dependency on w and adds a new dependency on $\chi=(\tau, \delta)$, where $\chi \in D$, then a new arc $e=((t, d), (\tau, \delta))$ appears in E of L. Thus E

contains the extension of a binary relation on D, and this relation expresses the intrinsic information structure of D.

Lexical stability and information structure can be factored independently in the evaluation of a dictionary. Consider lexical graph (4) from Figures 1 and 2, presented here in two views, one with node labels and one with lexical stability measurements:

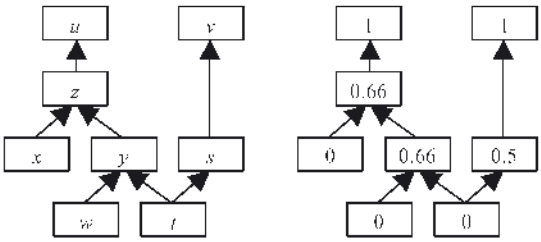

Figure 3. Two Views of a Simple Lexical Graph

Suppose that node z were changed such that it no longer depends on u. This modification changes the lexical stability values of nodes u and z to 0 and 1, respectively. No other stability values change. However, since z no longer depends on u, the transitive closure of lexical dependencies by which x, y, w, and t had previously related to u is disrupted. Because of the change in z, these nodes no longer base their definienda in whole or in part on u. Clearly the information structures of x, y, w, and t have changed, although their lexical stability measurements remain fixed. Since information structure can change independently from lexical stability, information structure and lexical stability must be factored independently in lexical graphs.

Except for orphans in a lexical graph, the meaning of a dictionary element is partially determined by its information structure. Thus a modification of v, $v \in D$ of lexical graph L, entails the possibility of change in meaning for v as well as for any other node u, where u is an element of the subgraph of L rooted at v. This is true because the information structure of u grounds its meaning in the information structure of v. Should v be positioned strategically within the global information structure of L–for instance, as the single root of L–then modification of v may transform the meaning of every term under definition in the dictionary.[5]

3. Discriminating between Dictionary Elements

For convenience, some device is required to express the foregoing notion of strategic positioning within a lexical graph. Given some node v, such a device will identify an aggregate stability value for v and the set of nodes in L sensitive to the meaning of v.

[5] As a practical matter, since L is a specialization of a DAG, any cyclical definitions in D of L must have been ameliorated prior to its creation. This treatment ignores terminological cycles, which arise when a concept is defined by direct or indirect reference to itself. For a formal treatment of terminological cycles, see [Ne91].

3.1 Aggregate Stability

Since lexical graph $L=(D, E)$ is a specialization of a DAG, it follows that the subgraph of L rooted at v is a lexical graph $L'=(D', E')$ where $v \in D$, $D' \subseteq D$, $E' \subseteq E$, and D' is the set of nodes from which v is reachable. Node v of L is reachable from node u if $v = u$ or L contains a path from u to v. Formally, v is reachable from u if and only if v is identical to u, v is adjacent to u, or there is some set of arcs $E' \subseteq ((u, x_i), (x_i, x_{i+1}), \ldots, (x_n, v))$, $E' \subseteq E$, where x_i and $x_i + 1$ are distinct and adjacent for $i=(0, \ldots, n)$. Thus for each node v in L, it is possible to isolate the contribution of the subgraph rooted at v to the global stability of L. Let this be known as $S[v]$, the aggregate stability of v:

Definition 2: Let $L'=(D', E')$ be the subgraph of L rooted at v. Then $S[v] = \sum_{x \subset D'} S(x)$.

The aggregate stability property of nodes in a lexical graph can serve as a crude comparator discriminating between nodes on the basis of the graph's global stability and its sensitivity to the information structure of each node under comparison. Where $S[v]>S[v']$, the global information structure of L is more sensitive to modification of v than v'.

Multiple roots are common in lexical graphs. Definition 2 isolates a partially ordered set $(D, S[v])$ that can serve to identify the aggregate stability of the root of some subgraph of L. Further, it might be applied as a comparator between subgraphs, thus identifying an aggregate stability value for each subgraph of a multiply rooted lexical graph.

3.2 Global Stability

The global stability of a lexical graph, $GS(L)$, can be defined as the sum of all $S(v)$ in L:

Definition 3: Let $L=(D, E)$ be a lexical graph. Then $GS(L) = \sum_{v \subset D} S(v)$.

Global stability may, for instance, be employed to characterize lexical graphs on the basis of the inherent integration of information concepts within their respective dictionaries. Where D of two lexical graphs is equivalent, $GS(L)$ is greater for the lexical graph with the greater interdependency between dictionary definitions. A similar comparator for arbitrarily selected lexical graphs is the relative global stability of each graph L, $GS_R(L)$, which expresses the relative integration of information concepts within any lexical graph:

Definition 4: Let $GS(L)$ be the global stability of L. Then $GS_R(L) \quad \frac{GS(L)}{D}$.

Where $GS_R(L)>GS_R(L')$, the information structure of L encodes a greater interdependence of information concepts than that of L'. As a practical matter, whenever $GS_R(L)$ is

relatively high or low, dictionary D of L will demonstrate a commensurately high or low level of conceptual integration.[6]

3.3 Fractional Stability

Summarizing briefly, Definition 2 demonstrates that for each node v in L it is possible to isolate the contribution of the subgraph rooted at v to the global stability of L. Using this technique, an aggregate stability value can be identified for an arbitrary subgraph of L. Definition 3, on the other hand, encapsulates in a single property the local lexical stability values distributed in a lexical graph. When construed in concert, these values can be used to express the fraction of a lexical graph's global stability contributed by the information structure of some discrete node. Let this property be known as the fractional stability of node v, $FS(v)$:

Definition 5: Let L' be the subgraph of L rooted at v. Then if $d(v)$ 0, $FS(v)$ 0; otherwise,

$$FS(v) = \frac{S[v]}{GS(L')}.$$

Fractional stability provides an insightful discriminator for the information concepts encoded in a dictionary. It imposes a weak partial order on the elements of D, identifying for each node a value in the interval between 0 and 1. For illustration, consider the two views of a lexical graph displayed in Figure 4:

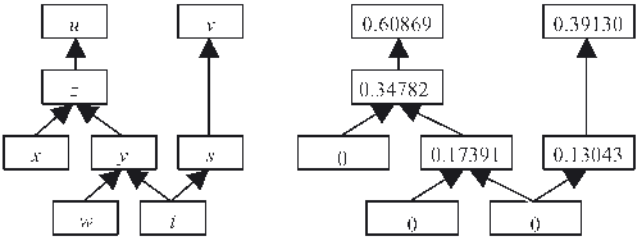

Figure 4: Fractional Stability in a Simple Lexical Graph

This is the graph from Figure 3 with fractional stability values labeling its nodes. It indicates how a greater $FS(v)$ value for a node correlates with a greater potential for transformation of meaning of dictionary elements in the lexical graph.[7] In this case, the fractional stability relation defines a partially ordered set $(D, FS(v))=\{t, w, x, s, y, z, v, u\}$. Meaningful change to a dictionary element at the lower end of the scale has little or no ramification for the lexical graph or the meaning of other dictionary elements. Conversely, modification of a dictionary element at the high end of the scale has

[6] Dictionaries with highly interdependent lexis tend to be rigid, narrowly domain-specific, and difficult to maintain. Even so, interdependence is necessary if a dictionary is to be useful and coherent. Thus some forms of dependency may be desirable, and others undesirable.

[7] High precision in fractional stability metrics is often necessary. The author has personal experience with mature data dictionaries from the aerospace and defense industries that exhibit nodes with incoming degree as high as 175, or outgoing degree as high as 14.

potentially extensive consequences for the configuration of the lexical graph and the meaning of other dictionary elements.

The fractional stability relation exposes this lexical graph property to measurement, enabling a scalar distribution of the information concepts encoded in a dictionary. Dictionary elements arranged at the lower end of the scale have information structures with few or no dependents. Meaningful change to their semantics can be introduced without effect on the remainder of the dictionary. In this sense, elements at the lower end of the scale are peripheral information concepts with respect to the global information structure of the dictionary. As for the information concepts arranged at the higher end of the scale, meaningful change to their semantics will be broadly propagated throughout the information structure of the dictionary. Hence, these elements encode the core information concepts of the dictionary. The chart displayed here depicts the elements of this partially ordered set for the simple lexical graph of Figure 4, along with the aggregate and fractional stability properties of each term under definition in the dictionary of that lexical graph.

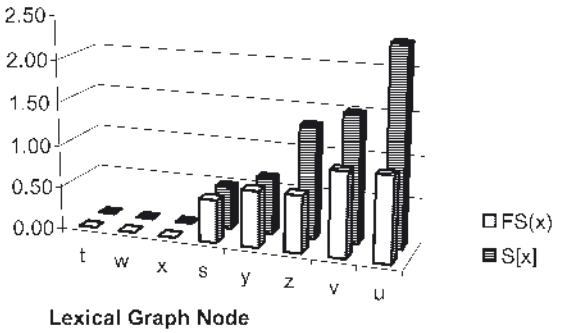

Figure 5: Some Properties of a Simple Lexical Graph

4.0 Conclusions

This paper has extended graph theory to encompass a new data configuration called lexical graphs. The objective has been to sketch a mathematical model to support the analysis of the form and content of lexicographic dependency information, whether it be contained in a conventional dictionary, a data dictionary, the semantic repository of a large business enterprise, or some other form of managed terminology.

Lexical graphs provide an insightful formalism for modeling the information structure of dictionaries. Lexical graph theory prescribes formal techniques for inducing a topology on a dictionary, measuring its lexical stability locally and globally, and distinguishing between core and peripheral information concepts encoded within it. The sorted topology of a dictionary is useful because it provides a learning sequence for the subject matter compiled in it, and lexical stability identifies the contribution of a particular

lexical entry to the structural integrity of the dictionary as a whole. Exploiting these properties, it is possible to identify a lexical stability value for a single dictionary entry, a subset of interdependent dictionary elements, and the dictionary as a whole. Since many algorithms practiced in computational linguistics resort to lexicographic data of one form or another, the theory of lexical graphs may benefit natural language engineering technologies of various kinds, from information-based approaches to frameworks for conventional natural language understanding.

ACKNOWLEDGEMENTS

The author is indebted to Ping Xue for useful discussions of the issues treated in this paper. This research has been supported in part by the Deepwater program of the United States Coast Guard.

Bibliography

[Kn68] Knuth, D. Semantics of Context-Free Languages. *Mathematical Systems Theory 2*, 1968, pp. 127-45.

[Ne91] Nebel, B. Terminological Cycles: Semantics and Computational Properties. In Sowa, J. Ed. *Principles of Semantic Networks; Explorations in the Representation of Knowledge*. Morgan Kaufmann, San Mateo, 1991; pp. 331-61.

[Wa72] Wall, R. *Introduction to Mathematical Linguistics*. Prentice-Hall, Englewood Cliffs, N.J., 1972.