Detecting Causally Embedded Structures Using an Evolutionary Algorithm

Chen Li
Department of Linguistics
University of Illinois at U-C
707 S Mathews Avenue,
Urbana, IL 61801
chenli@illinois.edu

C. Roxana Girju
Department of Linguistics
University of Illinois at U-C
707 S Mathews Avenue,
Urbana, IL 61801
girju@illinois.edu

Abstract

Causality is an important relation among events and entities. Embedded causal structures represent an important class, expressing complex causal chains; but they are traditionally difficult to uncover automatically. In this paper we propose a method for the efficient identification and extraction of embedded causal relations with minimal supervision, by combining a representation of structured language data with modified *prototype* theory specifically suited to the data type. We then utilize a form of *genetic algorithm* specifically adapted for our purpose to locate the likely candidate linguistic structures that contain causal chains. With this procedure, we were able to identify many embedded structures with complex causal chains in two corpora of different genres, applying this algorithm as a ranking procedure for all structures in the data. We obtained 79.5% percision for top quantiles of both of our datasets (BNC & novels).

1 Embedded Causality

Long chains of causal relations are frequently denoted by a complex embedding of multiple clauses through lexico-syntactic structures, structures which are causally linked. Following previous approaches (Menzies 2009, Beamer & Girju 2009), we define a causal relation as $e_1 \xrightarrow{cause} e_2$, where e_1 precedes e_2 temporally and, had e_1 failed to take place, e_2 would also not have taken place, or more generally, $P(e_2|e_1) > P(e_2|\neg e_1)$. This is a general and agreed upon definition of causality which encompasses various classes of causal types of interest (if one chooses to go deeper into this problem). Our unit of representation (for both the cause and the effect) is a *semantic frame*, given by a predicate and a list of arguments in the form $\phi(ARG_i, ARG_j, ARG_k, ...)$. This corresponds to a clause. Such clauses ocurring in embedded structures can form a causal chain. For example (from *Little Women*):

1. a smart shower at eleven had evidently quenched the enthusiasm of the young ladies who were to arrive at twelve for nobody came and at two the exhausted family sat down in a blaze of sunshine to consume the perishable portions of the feast (prepared in anticipation of the guests) that nothing might be lost (Alcott, 1868)

   (a) a smart shower at eleven had evidently quenched the enthusiasm of the young ladies who were to arrive at twelve
   (b) $\xrightarrow{cause}$ nobody came
   (c) $\xrightarrow{cause}$ the exhausted family sat down in a blaze of sunshire
   (d) $\xrightarrow{cause}$ consume the perishable portions of the feast
   (e) $\xrightarrow{cause}$ nothing might be lost

In this paper we focus on causal relations between clauses (marked or not by discourse markers).

1.1 Distinct characteristics

Each embedded causal structure has a *causer* entity identified by the main clause, and an effect event identified by the embedded (i.e. subordinate) clause. A class of semantically rich verbs is often present, that convey some notion of causation, coloring the causing event with additional *manner* of causation – verbs such as *inspire*, *suggest*, *prompt*, *bribe*, *incite*, *bully*, *force*, *compel*, etc. We call this class $\mathcal{MCC}$-verbs. Other verbs such as *cause*, *bring-about*, however, are just simple causatives (Girju 2003). Depending on its complexity, there may be one or more intermediate clausal structures that represent links in the causal chain, along with intermediate causal agents whose presence could have little specific semantic information, e.g. *"...caused the circumstances to line up in such a way as to..."*, but informs of its properties as a causal chain.

Due to the complexity of these elements and the intervening structures, there are many combinatorial possibilities, and the depths of such structures are potentially unbounded. So rather than finding a comprehensive set of *exemplars* that cover all cases, it is better to assemble patterns that represent a diffuse *prototype*, finding characteristic structures common in embedded causal frames, such as:

Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages 43–52,
Denver, Colorado, June 4, 2015.

i $ENTITY_{causer}$ ***caused it to come about*** *that* $ENTITY_{causee}$ $[PRED_{emb}$ ....]

ii $ENTITY_{causer}$ ***arranged*** *the events* ***so that it comes about*** *that* $ENTITY_{causee}$ $[PRED_{emb}$ ....]

iii $ENTITY_{causer}$ *had the forsight* ***to prepare the circumstances so that*** *it* ***comes about*** *that* $ENTITY_{causee}$ $[PRED_{emb}$....]

For all examples above, we can see that a subtree producing the terminals would be *"to come about that ...."*. A subtree like this can be used to further identify larger embedded structures as causal, and each embedded causative construction thus identified would contain one or more such subtrees.

1.2 Data

We considered two different genres: 1) the British National Corpus (BNC, 2007), and 2) novels from romantic fiction and historical novelas (mostly from Project Gutenberg, 2005), such as *The Great Gatzby*, *Pride and Prejudice*, *Little Women*, *Emma*, and *Lily of the Nile*. The training set consists of 500 positive instances (i.e., manually identified to contain a causal chain of at least one cause - effect relationship) which were selected from the 3^{rd} quarter of BNC. The testing sets consist of the 1^{st} quarter of BNC, and the novels set, respectively.

2 Previous work

There is a variety of approaches to causal relations in the literature, approaches which rely mostly on machine learning methods over high-dimensional semantic-feature spaces (Abe et. al., 2008; Berthard & Martin, 2008; Riaz & Girju; Do et. al., 2011; Radinsky et. al. 2012 / 2013; Oh et. al. 2013; Hashimoto et. al., 2014; etc). Other researchers have focused on pre-identified lexico-syntactic patterns (Khoo et. al. 2001; Girju 2003) which they use to bootstrap an Expectation-Maximization procedure (Chang & Choi 2006; Paul et. al. 2009) for causality and similar semantic relations. Furthermore, these parametric and pattern recognition works are generally fucussed on pair-wise causal relations between event representations. We instead focus on linguistic structures of unbounded complexity that are capable of expressing sequences of events involved in long causal chains. Our work explores novel representations of causality, and procedures rooted in evolutionary computing in order to deal with the structural complexity of these expressions as well as retain the flexibility of parametric approaches.

3 Diffuse prototype

We need to encompass available lexico-semantic (symbolic) and morphosyntactic (structural) information into a single representation that can be compared and transformed. And since our goal is to extract causal chains from complex structures, the representation needs to generalize the information over the member frames/clauses. We mostly focus on the intervening information and structural configuration between clausal subtrees, where the substructures are found based on sub-graph isomorphism between two positive samples treated as trees. The ideal product would be a set of maximally complex sub-structures in the reflection of their causality, which would not compromise their ability to generalize over all embedded causal structures. In this case, a purely parametric approach will not work for any tree structure of sufficient size, given the number of binary parameters that would need to represent the presence or absence of an edge $\langle v_i, v_j \rangle$ is $\mathcal{O}(n(T)^2)$. And thus, the number of possible configurations comes to $\mathcal{O}(2^{n(T)^2})$ without taking into account labels or other sources of complexity. For potential cognitive models of categorization, *prototype* and *exemplar*s are the primary theories most frequently considered. A single prototype is ideal for representing a set of similar objects that can be unimodally represented in feature space. A set of *exemplar*s has the advantage of allowing distributions in many modes in feature-space, each cluster being represented by a single exemplar.

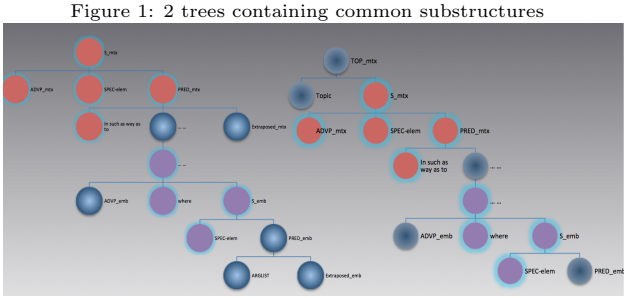
Thus, we propose and formulate a novel categorial model combining strengths of both prototype and exemplars, with a graph theoretic focus. Like *prototype*, it provides few structures far more concise than sample-set, allowing a high degree of generalization. Like *exemplar*s, it is adaptable in a multi-modal distribution over naturally defined feature space, with a wide coverage of subtypes. This we will term a *diffuse prototype* of the class, which are shared graph-theoretic substructures of at least two postitive samples. Given a feature space $X = [x_{[1]}, x_{[2]}, \ .... \ x_{[n]}] \in \{0,1\}^n$, a substructure, as a component within a *diffuse prototype* ($\mathcal{DP}$), is $X_s = \{x_{[\kappa_j]}\} \mid j \in \kappa \sqsubset [1, 2, ..., n]$ such that $\exists Y^p, Y^q \in Y \ \forall j \in \kappa \ [Y^p_{[\kappa_j]} = Y^q_{[\kappa_j]}]$, where Y = set of positive samples for that semantic class. Thus, the samples Y^p, Y^q agree on some substructure within the feature space. When the feature space is structured in some way, an additional constraint of *contiguity* is necessary. X_s above must follow *linear contiguity* for its contiguity definition. This requires that $\forall i, j \in \kappa \wedge X_{[\kappa_i]}, X_{[\kappa_j]} \in X_s$ and where $P_{i \rightarrow j} := \lhd i, \ ..., \ j \rhd$ is some consecutive sequence $\sqsubset \mathbb{N}$, we have that $\forall k \in P_{i \rightarrow j} \ [\kappa_k \in X_s]$ ($\sqsubset$ here symoblizes sub-sequence relation). So now in this example, any substructure must be restricted by some linearly contiguous region of X. We have a more complex structure of the feature space - the notion of *continguity*, referred to as $N^+_T(v_i)$ and $N^-_T(v_i)$, with which we determine the allowable extensions of

any substructure T_s. The allowed substructures are:

$$T_t = G(X_s) \begin{cases} \forall v_p, v_q \in V(T_t) \\ \exists P_{p \to q} := \ \lhd i, \ ..., k, ... \ j \rhd \ \sqsubset \kappa^P \\ \forall \kappa^P_k, \kappa^P_{k+1} \in \kappa^P \left[v_{[\kappa^P_{k+1}]} \in N^+_T(v_{[\kappa^P_k]}) \right] \end{cases} \tag{1}$$

Where κ^P is a specific ordering of $V(T)$ that conforms to the path P. The only types of X_s sought are those that form a proper subtree T_t of the original T. This is a natural way to allow generalization into members of the $\mathcal{DP}$, and thus some fragmented forest subgraph of T is not desirable. As an illustration, the trees T and T' in Figure 1 contain a pair of substructures T_s and T_t corresponding to red / violet regions. The shared subgraphs are used to find

Figure 1: 2 trees containing common substructures



yet some other T'' where variable (blue-grey) regions differ from either T or T'.

4 Extraction procedure

The key difficulty is the *isomorphic* comparison of two trees. For this, we developed a form of graph-theoretic genetic algorithm, simulating the growth of subtrees shared between two reference trees T, T'.

4.1 Baseline genetic algorithm

Inspired by *On the Origin of Species* (Darwin 1859), *genetic algorithms* are a class of adaptive algorithms (Turing 1950; Barricelli 1962; Rechenberg 1973; Holland 1975; de Jong 1975), with wide array of application (Brindle 1981; Baker 1985 / 1989; Goldberg 1989; Goldberg & Deb 1993; Fogel 1998). Our algorithm has similarities to *genetic programming* (Cramer 1985; Schmidhuber 1987), and to aspects of the original biological model, beyond traditional GA, due to greater variability afforded by substructure growth. Two processes are responsible for growth and diversification of chromosomes, *mutation* and *recombination* in GA. An elimination stage culls a part of the population with regard to some notion of *fitness* (*directed selection*), its magnitude determined by *carrying capacity* (Goldberg et. al., 1991).

4.2 Proposed modifications to GA

We adapted the baseline GA, redefining the operators graph-theoretically according to specific structure types in the $\mathcal{DP}$. For our evolutionary algorithm, we have thouroughly reformulated the three primary operators, *non-homogenizing*, *homogenizing*, and *culling*, as well as how gene loci are structured, from the baseline GA. For a minimal ecological niche, we find some lexico-syntactic cues and associated structures discussed in Section 1.1 that is shared by at least a pair of positive samples. To better discriminate between non-causal structures and embedded causal structures, we need to maximize the complexity of the $\mathcal{DP}$ members, in order to minimize the number of possible $T \in \mathbb{T}$ (set of samples) that could contain such, thus maximizing the specificity of $\mathcal{DP}$.

4.2.1 Individual and population

Our genotype is cast as a piece of structural information within some induced subtree T_s of $\langle T, T' \rangle$ that conveys causality, so *'chromosome'* is modeled as the set of parameters necessary to encode T_s denoted as ξ^{T_s}. Thus, the phenotype is simply whether ξ^{T_s}, once decoded into T_s fits inside the ecological niche as induced subgraph. Whether a *"phenotype"* is well adapted for the *"ecological niche"* can simply be a subgraph isomorphism test, which hereon we will denote as $\mathcal{I}^{\mathcal{S}}(T_s, T)$. The baseline GA represents chromosome modeled an ordered set of traits with linear contiguity. Since it is highly inefficient to represent all structural information of a chromosome as individual binary parameters, we redesigned this as a graph-theoretic representation of the linguistic structure. The members of $\mathcal{DP}$ are represented as subgraphs within embedded causal structures, so each GA-operator must be reformulated according to graph-theoretic concepts.

We will use standard graph theoretic notations, where $G = \langle V, E \rangle$ (vertices and edges); and $N^{+/-}_T(v)$ is the operator that locates the neighbors set of $v \in V$ of tree T in the $+/-$ direction. The genetic makeup of an individual is modeled as a single chromosome ξ^{T_s}, so the entire set of such sub-structures of $\langle T, T' \rangle$ becomes our population. Following our definition of *diffuse prototype*, a chromosome is not an ordered set of parameters, but a configuration of subgraph; and location of gene loci is not its linear position, but its relative location WRT to others, in a tree structure.

$$\xi^{T_s} = \begin{cases} \langle v_r, v'_r \rangle \ \Big| \ \left[v_r \in V(T) \wedge \nexists v_s \in V(T_s)[v_s \in N^-_T(v_s)] \right] \\ \qquad \bigwedge \left[v'_r \in V(T') \wedge \nexists v'_s \in V(T'_s)[v'_s \in N^-_{T'}(v'_s)] \right] \\ \left\langle V_l = \left\{ v_l \Big| \exists v_m \in N^+_T(v_l)[v_m \notin N^+_{T_s}(v_l)] \ \vee \ |N^+_T(v_l)| = 0 \right\} \right. \\ \left. V'_l = \left\{ v'_l \Big| \exists v'_m \in N^+_{T'}(v'_l)[v'_m \notin N^+_{T_s}(v'_l)] \ \vee \ |N^+_{T'}(v'_l)| = 0 \right\} \right\rangle \end{cases} \tag{2}$$

ξ^{T_s} must contain locations of the boundary nodes of substructure within T; such boundaries of both T and T' are contained within ξ^{T_s}, where each point in the boundary is implemented as a pointer to a tree node. So ξ^{T_s} is a collection of pointers WRT $\langle T, T' \rangle$: By moving pointers around $V(T), V(T')$, we can decode T_s. The $\rho-$ and $\lambda-$operators indicate the *'root'* and *'leaves'* of T_s WRT any habitat tree $\check{T}$.

$$\begin{cases} v_r = \rho_{\check{T}}(\xi^{T_s}) \mid v_r \in V(\check{T}) \\ V_l = \lambda_{\check{T}}(\xi^{T_s}) \mid \forall v_l \in V_l [v_l \in V(\check{T})] \end{cases} \tag{3}$$

The initial generation $\mathcal{G}^0$ consists of identical single nodes between T, T', and G is max generation limit, $\{\rho_{\check{T}}(\xi^{T_s})\} = \lambda_{\check{T}(\xi^{T_s})} \wedge \{\rho'_{\check{T}}(\xi^{T_s})\} = \lambda'_{\check{T}}(\xi^{T_s})$.

4.2.2 Non-homogenizing operator

The non-homogenizing operator should be designed to create new gene variations in the population, this is equivalent to *mutation* in baseline GA. Since we cannot efficiently encode all possible subgraphs of T, we use the far more efficient ξ^{T_s}. So we reformulated the non-homogenizing operator as a process that grows a tree sub-structure one edge+node at a time, and thus introduces new degrees of freedom each generation. We define an operation that might add a new vertex $v_i \in V(T) \setminus V(T_s)$ and edge $\langle v_i, v_j \rangle$ or $\langle v_j, v_i \rangle \in E(T), v_j \in V(T_s)$. This is easiest realized in two subtypes, due to the directed nature of T and has no effect on the genetic makeup of the following generation. The previous configuration ξ^{T_s} remains if conditions are not met.

$$\mu_r(\xi^{T_s}, \langle T, T' \rangle) = \begin{cases} \Big\langle \langle v_j, v'_j \rangle, \langle \lambda_T(\xi^{T_s}), \lambda'_{T'}(\xi^{T_s}) \rangle \Big\rangle \\ \quad \Big| \ \Big(v_j = \rho_T(\xi^{T_s}),\ v_i \in N_T^-(v_j) \Big) \\ \quad \bigwedge \Big(v'_j = \rho'_{T'}(\xi^{T_s}),\ v_i \in N_{T'}^-(v_j) \Big) \\ \quad \bigwedge \Big(\varsigma(v_j) = \varsigma(v'_j) \Big) \end{cases} \tag{4}$$

where l, r are leaf and root directions. When both T, T' agree on a common addition to the current substructure, it returns $\xi^{T'_s}$, grown from the structure of T_s in the r-direction. Here is μ in the l direction (the operator $v_h =_{T_t} v_k$ denotes that they are topologically equivalent with respect to the substructure T_t that is shared within the pair $\langle T, T' \rangle$):

$$\mu_l(\xi^{T_s}, \langle T, T' \rangle) = \begin{cases} \Big\langle \langle \rho_T(\xi^{T_s}),\ \rho'_{T'}(\xi^{T_s}) \rangle, \langle \lambda_T(\xi^{T_s}) \cup \{v_i\}, \\ \lambda'_{T'}(\xi^{T_s}) \cup \{v'_i\} \rangle \Big\rangle \ \Big| \ \exists v_i \in V(T), \\ v'_i \in V(T'), v_i =_{T_s} v'_i \ \Big(v_i \notin V(T_s) \wedge \\ \exists v_j \in \lambda_T(\xi^{T_s}) \ [\ v_i \in N_T^+(v_j)\ ] \Big) \bigwedge \\ \Big(v'_i \notin V(T_s) \wedge \exists v_j \in \lambda'_{T'}(\xi^{T_s}) \\ [\ v'_i \in N_T^+(v_j)\ ] \Big) \bigwedge \Big(\varsigma(v_i) = \varsigma(v'_i) \Big) \end{cases} \tag{5}$$

During the non-homogenizing stage of a generation, each individual ξ^{T_s} within the population has a chance to undergo either $\mu_r(\xi^{T_s}, \langle T, T' \rangle)$ or $\mu_l(\xi^{T_s}, \langle T, T' \rangle)$. The probabilities are mediated by the random variables $R^{\blacktriangle}$ and $R^{\blacktriangledown}$; the ratio between $R^{\blacktriangle}, R^{\blacktriangledown}$ is governed by the mean branching factor of T, T', so to ensure even growth in all directions.

4.2.3 Homogenizing operator

A homogenizing operator in biological systems or GA randomizes the distribution of alleles and redistribute new allelic types among the population, by exchange of information between distinct units of inheritance between homologous loci; the most prevalent is *recombination*. For our purposes, this is similar to individual haploid organisms exchanging genetic material (e.g. plasmids). So we reformulate the homogenizing operator as process of separating and regrafting tree substructures together to form new tree configurations. These disparate configuration types are analogous to single-point and 2-point cross-overs in linear genomes. The former is a single contiguous region of shared loci between T_s, T_t, $\varkappa_{\diamondsuit}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$. One or more significant regions being shared between to substructures gives us a good anchor for performing cross-over of the remaining, differentiating regions of the substructure, essentially a form of *elitism* (Chakraborty & Chaudhuri, 2003; Mashohor, 2005; Yang, 2007; Chudasama, 2011; Yaman & Yilmaz, 2012; Bora et. al. 2012; etc). We do so by identifying the regions of two substructure with identical graph topology as well as labels of each's vertices. Given some minimum size for the shared region $c_{\diamondsuit}$:

$$\varkappa_{\diamondsuit}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t} \begin{cases} V_m(T_s) \ \Big| \ \psi^{T_s,T_t}(V_m), |V_m| \geq c^{\diamondsuit} \\ \quad \nexists V'_m \subset V(T_s)\ [\psi^{T_s,T_t}(V'_m) \wedge |V'_m| > |V_m|]\ ; \\ \quad \psi^{T_s,T_t}(V_p) = \Big(\ \forall v_i \in V_m [\varsigma(v_i) = \varsigma(v'_i)] \ \Big) \\ \quad \bigwedge \Big(\langle v_i, v_j \rangle \in V_p(T_s) \leftrightarrow \langle v'_i, v'_j \rangle \in V_p(T_t) \Big) \end{cases} \tag{6}$$

The second type is two discontiguous regions of shared loci between T_s, T_t; denoted as $\varkappa_{\bowtie}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, a pair of disjoint maximum common subgraphs of T_s, T_t. This is analogous to the previous formulation for the shared region of single-point crossover scenario, except that there are two discontiguous regions with a differentiating graph region in between. where we may denote the elements in the pair as $\varkappa_{\bowtie\ [s]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, & $\varkappa_{\bowtie\ [t]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$. These shared regions of T_s, T_t of $\varkappa^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$ essentially function as a highly specialized form of *rank elitism* (Chakraborty & Chaudhuri, 2003; Mashohor, 2005; Yang, 2007; Chudasama, 2011; Yaman & Yilmaz, 2012; Bora et. al. 2012; etc), that operates specifically with our sub-structures, where the $\varkappa-$regions function to filter pairs of $\langle T_s, T_t \rangle$ so only the highly compatible pairs would undergo homogenization. We denote subgraph relation as $\trianglelefteq$, a set of all connected componts of G as $\circledast(G)$. We use $\binom{S}{c}$ for denoting the choosing of c elements from the set S, and employ a random variable R^S, so $\binom{S}{c}^{R^S}$ preferentially chooses those of the greatest size. Here, $\varrho_{\diamondsuit}$, the differentiating regions that may be grafted

onto another corresponding substructure, by finding the set of disjoint graph regions ($\divideontimes$) of a induced subgraph of the substructures T_s, T_t, induced with vertices outside of the shared ($\varkappa_\Diamond -$) region.

$$\varrho_{\Diamond}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t} \begin{cases} \left\langle \binom{S}{2}^{R^S}, \binom{S'}{2}^{R^S} \right\rangle \;\Big|\; \\ S = \divideontimes\Big(\phi\Big(\; (V(T_s) \setminus V(\varkappa_{\Diamond}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}))(T_s) \;\Big)\Big)\,, \\ S' = \divideontimes\Big(\phi\Big(\; (V(T_t) \setminus V(\varkappa_{\Diamond}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}))(T_t) \;\Big)\Big)\,; \\ \phi(E_p) = G(\; \{v_q | \langle v_q, v_r\rangle \in E_p \vee \langle v_r, v_q\rangle \in E_p\},\; E_p\;) \end{cases} \tag{7}$$

We denote the elements within the target loci range: $\varrho_{\Diamond\ [S,0]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, $\varrho_{\Diamond\ [S,1]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, $\varrho_{\Diamond\ [T,0]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, $\varrho_{\Diamond\ [T,1]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$. The corresponding $\bowtie$ type regions is obtained by locating the disjoint regions of the induced subgraph of T_s, T_t, induced by the vertices outside of the shared $\varkappa_{\bowtie}$ region:

$$\varrho_{\bowtie}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t} \begin{cases} \Big\langle T_u,\; T_v \Big\rangle \;\Big|\; \psi(T_u, S, T_s) \;\wedge\; \psi(T_v, S', T_t) \\ S = \divideontimes\Big(\phi\Big(\; (V(T_s) \setminus V(\varkappa_{\bowtie\ [s]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}))(T_s) \;\Big)\Big)\,, \\ S' = \divideontimes\Big(\phi\Big(\; (V(T_t) \setminus V(\varkappa_{\bowtie\ [t]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}))(T_t) \;\Big)\Big)\,; \\ \phi(E_p) = G(\{v_q | \langle v_q, v_r\rangle \in E_p \vee \langle v_r, v_q\rangle \in E_p\},\; E_p\;); \\ \psi(T_w, S_x, T) = T_w \in S_x \wedge \exists v_i, v_j \in \varkappa_{\bowtie}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t} \Big[\exists P_{i,j} = \\ [v_i, ..., v_j] \trianglelefteq T, [\exists \langle v_h, v_k\rangle \in E(T_w)\; \langle v_h, v_k\rangle \in E(P_{i,j})]\Big] \end{cases} \tag{8}$$

Random variables $R^{\Diamond}$ and $R^{\bowtie}$ give the probabilities of each $\Diamond$ or $\bowtie$ type operator would be applied. $\Diamond$-type operation is shown for the [1]−component ([0]−component would be analogous), as $\eta_{\Diamond}^{s \rightarrowtail t}(s \rightarrowtail t)$ (t [1]−component grafted onto T_s), and $\eta_{\Diamond}^{t \rightarrowtail s}(T_s, T_t)$ (s [1]−component grated onto T_t). $\phi_{\Diamond}(\cdot)$ and $\psi_{\Diamond}(\cdot)$ provide configuration of the nodes' relations WRT the ϱ and $\varkappa$ regions, omitted due to space constraints.

$$\eta_{\Diamond}^{s \rightarrowtail t}(T_s, T_t) \begin{cases} V^{s \rightarrowtail t} = (V(T_s) \setminus V(\varrho_{\Diamond\ [S,1]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t})) \;\cup\; V(\varrho_{\Diamond\ [T,1]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}) \\ E^{s \rightarrowtail t} = E(\; V^{s \rightarrowtail t}(T_s)\;) \cup E\Big(\varrho_{\Diamond\ [T,1]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}\Big) \cup \{\langle v_i, v'_j\rangle\} \\ \Big|\; \Big(\phi_{\Diamond}(\langle v_i, v_j\rangle) \vee \phi_{\Diamond}(\langle v_j, v_i\rangle)\Big) \\ \bigwedge \Big(\psi_{\Diamond}(\langle v'_i, v'_j\rangle) \vee \psi_{\Diamond}(\langle v'_j, v'_i\rangle)\Big)\,; \end{cases} \tag{9}$$

It takes the necessary vertices from the graft [1]−component of T_t, and the remainder of T_s, and add a new edge so that they are still attached in the same configuration as they were in T_s and T_t. The $t \rightarrowtail s$ process is the mirror image of $s \rightarrowtail t$ when the same $\varrho_{\Diamond}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t} \Leftrightarrow \varkappa_{\Diamond}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}$, omitted due to space. Correspondingly, the $\bowtie$ −type operation is similar to a two-point cross-over, with two new edges $\langle v_i, v'_j\rangle$, $\langle v_p, v'_q\rangle$ necessary for the new composite form. We define $\bowtie$ type operation, with the two recombinations as $\eta_{\bowtie}(T_s, T_t)$, and demonstrate one direction of grafting of component onto the remainder of T_s, the opposite direction is analogous.

$$\eta_{\bowtie}(T_s, T_t) \begin{cases} V^{\bowtie} = (V(T_s) \setminus V(\varrho_{\bowtie\ [S]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t})) \;\cup\; V(\varrho_{\bowtie\ [T]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}) \\ E^{\bowtie} = E(V^{\bowtie}(T_s)) \cup E(\varrho_{\bowtie[T]}^{T_s \overset{\mathcal{I}}{\longleftrightarrow} T_t}) \cup \{\langle v_i, v'_j\rangle, \langle v'_p, v_q\rangle\} \\ \Big|\; v_i \neq v_q \;\bigwedge \Big(\Big(\phi_{\bowtie}(\langle v_i, v_j\rangle) \vee \phi_{\bowtie}(\langle v_j, v_i\rangle)\Big) \bigwedge \\ \Big(\phi_{\bowtie}(\langle v_p, v_q\rangle) \vee \phi_{\bowtie}(\langle v_q, v_p\rangle)\Big)\Big) \bigwedge \Big(\Big(\psi_{\bowtie}(\langle v'_i, v'_j\rangle) \vee \\ \psi_{\bowtie}(\langle v'_j, v'_i\rangle)\Big) \bigwedge \Big(\psi_{\bowtie}(\langle v'_p, v'_q\rangle) \vee \psi_{\bowtie}(\langle v'_q, v'_p\rangle)\Big)\Big) \end{cases} \tag{10}$$

Similar to $\Diamond$, the $\bowtie$ −type operator takes the necessary nodes from the two shared regions between T_s, T_t and any non-shared regions not between shared regions. It then locates the nodes in the graft component and edges between them. Finally, it includes two new edges, making the connection between graft and host, while preserving the local configurations at the attachment points. It is easier illustrated pictorially, such as $\Diamond$−type operation in Figures 2 - 3; the single red region are shared between the substrutures, while the orange, green regions within T_s, and yellow, purple regions in T_t, undergo re-grafting into new host structures from one figure to the next.

Figure 2: before single-point crossover

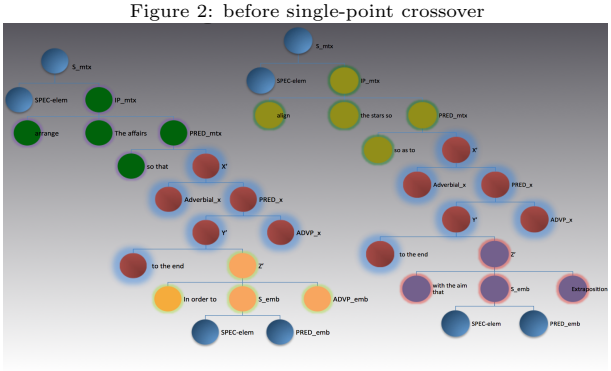
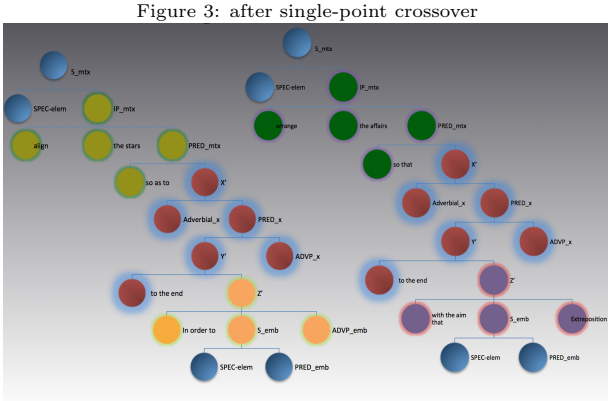


Figure 3: after single-point crossover



4.2.4 Culling operator & genetic drift

Death is an essential component of evolution in nature; with significant death rate, natural selection has an opportunity to apply its pressure. In a biological system, this process is a mixture of *directed selection*, (depends on the fitness of an organism in an ecological niche), or migration patterns among niches; and some *randomization* in selection, which in nature include *genetic drift* and *immigration/emmigration*. Directed selection is the primary driver for adaptation, when the environment remains static over several generations. The primary metric of usefulness of any T_s is its complexity measured as $n(T_s)$, which entails the maximization of the number of potential non-homogenizing operations on T_s^l. So the base formulation of *fitness* is based the total capacity to reproduce (potential rate * reproductive span), termed *fecundity*, and the actual rate given population and environmental factors, termed *fertility*. This important ratio is $f(T_s) = \frac{fertility}{fecundity} = \frac{\mathfrak{f}^{T_s}}{\mathfrak{e}^{T_s}}$. Also a factor in the usefulness of sub-structure T_s is the distribution of terminal symbols of T_s within the corpus. We incorporated *lift* of tokens of tree terminals within positive sample, against all tokens in the training data. Let $\tau(T_s)$ be a function linearizes the terminals of the tree T_s, and where X_E is the set of terminal sequences from the positive samples, and $X_{E\&i}$ are samples showing both traits; the fitness F is:

$$F(T_s) = \begin{cases} \sum \mathcal{L}(X_E \Longrightarrow x_j) \\ f(T_s) \propto \frac{\mathfrak{f}^{T_s}}{\mathfrak{e}^{T_s}} \cdot \frac{x_j \in \tau(T_s)}{|\tau(T_s)|} \\ \mathcal{L}(X_E \Longrightarrow x_j) = \frac{\mathcal{S}(X_{E\&j})}{\mathcal{S}(X_E) \times \mathcal{S}(X_i)} \mid x_j \in X_i \\ \sum \mathcal{N}_j \\ \mathcal{S}(X_i) = \frac{x_j \in X_i}{\sum_{x_k \in X} \mathcal{N}_k} \;\Big|\; \mathcal{N}_j \propto n(X_j) \end{cases} \tag{11}$$

Since degrees of freedom increase over generations, Boltsmann selection is unnecessary and may even delay arrival at global maximum. The procedure used is a *roulette selection*, since variability of fitness within a single generation is small. *Genetic drift* is not an issue here, since the degree of freedom available increases with each generation's non-homogenizing operation. Each testing sample is tested against the extracted substructures. This process still has high time complexity, potentially $\mathcal{O}(n^{k+4.5})$ (k is the degree limit) (Bodlaender, 1988). Additional pre-filters (i.e., number of vertices, degree-list, label-histogram of V) are applied to further reduce complexity.

5 Test results

5.1 Dataset and model parameters

The BNC is a mixed corpus with complex genres such as parliamentary proceedings and news articles. The training set needs to have sufficiently complex frames to have a significant probability of being embedded causals. Other non-training parts of BNC, as well as the novels corpus, were used for testing. The labelled data is lexed and parsed, and some tree transformations are detected and reconstructed, and separated into semantic frames. The BNC-testing data contained 196314 lines, and novels set 129695 lines. For the **novels** testing set, 26356 instances of semantic frames were detected, and for the **BNC** testing set, 31807 instances of frames were detected. This procedure provides no specific threshold, since it is not binary, but produces a score for likelihood of complex causality. For evaluation, due to output frame count, and the fact that embedded causal structures are a small fraction of all possible clauses, standard precision + recall over the corpus is not feasible. The most sensible method is a sparse *quantile*-based annotation. We annotated three sets of $k = 100+$ (actually 115 each, to ensure at least 100 determinable). The annotation of this initial testing phase was performed by one of the authors. The labels for sample are Y (causal), N (non-causal), and U (undeterminable)

5.2 Ranking evaluation

The results are ranked sets of samples. A positive causal chain sample will contain at least some clearly identifiable $e_i \xrightarrow{cause} e_j$, where e_i, e_j are events expressed by clauses in the surface sequence. It is not required that each pair of adjacent pair of events e_i, e_{i+1} would be causal; and some causal relation $\langle e_i \xrightarrow{caus} e_j \rangle$ may not be immediately adjacent pairs (may skip some event in sequence). We explored how quickly the result by annotating the next several quantiles, each with the aforementioned approximately 115 samples to guarantee each quantile having at least 100 determinable ones. Since it is very labor intensive, we annotated each until a trend in quantile precision emerges, which was 7 for BNC-testing, and 10 for novels, when we observe a large difference from the highest quantile and a trend tending to a distribution tail. There are now 805 annotated samples from top 7 quantiles (Y:225, N:457, U:93) from the BNC-testing, where the top quantile had precision of 0.795, next highest quantile 0.677, and 7th quantile 0.133 (quantile precision is shown in Figure 4). There is a total of 1150 from top 10 quantiles (Y:401, N:672, U:77) from novels; the top quantile had a precision of 0.800, next highest 0.574, 7th quantile 0.206 (shown in Figure 5). The top 2 quantiles for each set is significantly above %50, thus a relatively high-confidence threshold can be set at top 200 for a binary classification task. The samples mostly contain 2-5 events in a causal chain, with the longest of 7.

Figure 4: BNC precision in its top 7 absolute quantiles

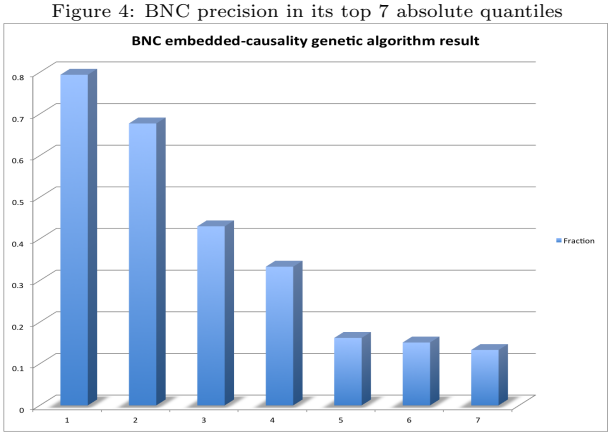
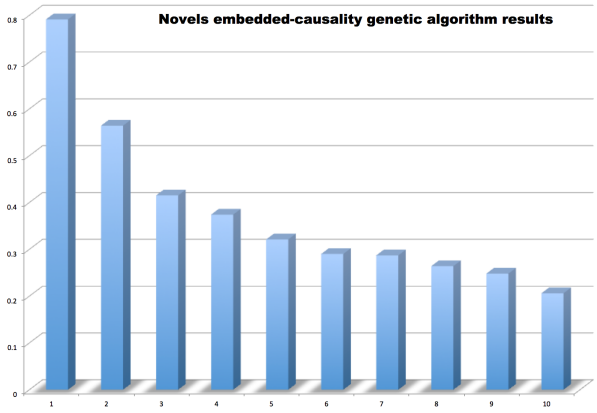


Figure 5: Novels precision in their top 10 absolute quantiles

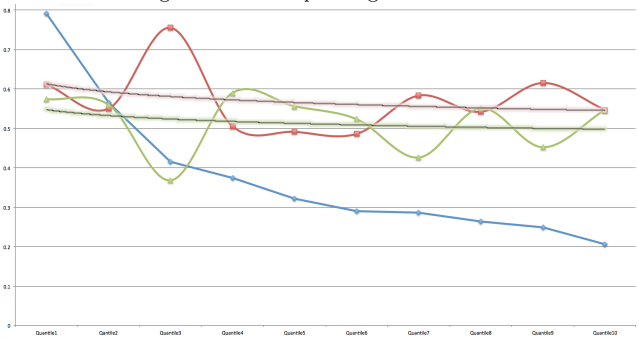


5.3 Comparison with baselines

We compared the results of our system to baselines, a textual entailment system as well as an n-gram model; since annotation is highly labor intensive, the annotated data are from the top 10 quantiles of our ranking. Thus these samples are already pre-selected by our system to be relatively likely to be causal; so we mainly test to see if a correlation with our system exists, and whether they produce the same gradient of precisions that rank from highest quantile to the lowest among these 1150 samples. For each, we expect some positive correlation with ours; but our system, being more specifically designed for complex causality, should outperform each.

We are unaware of any comparable system for complex causality, so textual-entailment (TE) is the most similar to our task. Thus, we used the TE system VENSES (Delmonte et. al. 2007/2009). This test is not approrpiate for the original purpose of VENSES, but is done with our data and annotation to see any correlation to our results, a comparison of the closest system. For any given sample of testing set, we determine whether any pair of the multiple clauses, is identified as entailed by VENSES. We compared the results against our gold standard (for embedded causality). The samples are the top 10 quantiles of the novels data-set (set with the most annotated samples), ranked according to our algorithm. Figure 6 contain the TE fraction of each quantile according to VENSES (red), whether VENSES judgement on TE is consistent with our human annotation on causality (green), and our system's output (blue). TE results

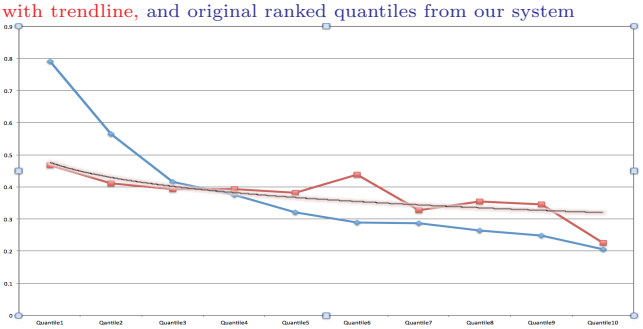
Figure 6: fractions of TEs according to VENSES, fraction of VENSES Y/N output same as human judgement on causality, and our system; for each of the top 10 quntiles for *novels*. The black lines with shading are the corresponding trend lines



labels contained many false negatives, since it is not designed for causality. This also serves as a baseline for our system, given TE is the closest system available for testing, where our system overperformed significantly given the task of complex causality.

Causal chains are highly sequential structures, so an n-gram model is a reasonable method for comparison. We also produced a standard n-gram model with smoothing and back-off, trained on the same training data as our system. Each sample of multiple clauses/frames is presented a a single sequence of terminal tokens. We determined that a trigram model is the optimum to obtain good specificity and avoid over-training. Thus, we tested it against each of the annotated testing samples, and produced a ranked score using the harmonic mean of probability of each token in the sequence according to the trigram model. Given that the testing samples are preselected by our system to be top-10 quantile, the n-gram model provides a re-ranking of these. We examined this re-ranking to see whether we get the same differentiation in precision in the new 10-quantiles of the same size after re-ranking (Figure 7). Thus the results of our system are also weakly correlated with n-gram re-ranking; but our system provides much better Y/N separation of the gold-standard in the trajectory over the top quantiles, and provides a more consistent and monotonic trend.

Figure 7: precision in re-ranked quantiles according to n-gram, with trendline, and original ranked quantiles from our system



5.4 Further Analysis

Given *causality* has many divergent definitions, we used a detailed characterization scheme allowing each annotator to select from "categories" of causations. Each of these characterizes one frequently accepted aspect of causation, including the four classical *'material' (constitution of component sub-events)*, *'formal'*, *'efficent'*, and *'final' (purpose)* causes (Aristotle 350 B.C. / 322 B.C.), which are often regarded in cognative studies as relevant aspects of causation that humans use in recognition of causality (Rachlin 1992, Hogan 1994, Killeen 2001, Killeen & Nash 2003, Alvarez 2009); as well as other common aspects of causality, *'cause of necessity (enablement)'*, *'cause with intermidate volition' (inducement)'*, and *'latent causal chain (outcome)'*. We also labelled the top 150 samples of the novels set, for the presence/absence of each of these 7 classes. Since long causal chains may contain multiple relations of different semantic types in one sequence, a sample may have multiple labels. The number and percent of the top 150 ranked samples are ❶ *efficient:* 17, 11.3%; ❷ *necessity:* 36, 24.0%; ❸ *formal:* 42, 28.0%; ❹ *final:* 40, 26.7%; ❺ *inducement:* 44, 29.3%; ❻ *material:* 17, 11.3%; ❼ *latent:* 10, 6.7%; which has a wide distribution among the 7, and has no particular dominant class. It is unsurprising that *latent causal chain* is contained in the least number of samples, since it is also the most difficult for people to detect. We here provide some top-quantile samples with the said annotation with a variety of classification scheme labels:

- eurotunnel is already in default of its credit agreement with the bank synidcate, [that it] is seeking an extra xx billions on top of the xx billions raise so far .
  eurotunnel is already in default of its credit agreement with the bank syndicate $\xrightarrow{efficient}$ *it is seeking an extra xx billions on top of the xx billions raised so far*

- before the housewives could rest several people called and there was a scramble to get ready to see them (receive them with hospitality)
  several people called [the housewives to visit] $\xrightarrow{efficient}$ *there was a scramble to get ready* $\xrightarrow{purpose}$ *to see them* (here meaning receiving the guests)

- she tries to find highborn women to bear him a son that she can take in as her own
  she tries to find highborn women $\xrightarrow{enables}$ *to bear him a son* $\xrightarrow{enables}$ *she can take in as her own*

- by late afternoon, I (Cleopatra Selene II) joined the rest of the women of the household Lady Octavia took it upon herself to [Lady Octavia] teach me (Cleopatra Selene II) to spin whorl
  I joined the rest of the women of the household $\xrightarrow{constitute}$ *Lady Octavia took it upon herself* $\xrightarrow{purpose}$ *teach me* $\xrightarrow{purpose}$ *spin wool*

- I (Cleopatra Selene II) was a Ptolemy princess (meaning descended from Hellenic-pharonic bloodline), a queen in exile who must bide her time until she could think of some plot, some plan to [some plot/plan] return her to her throne
  I was a Ptolemy princess $\xrightarrow{constitute}$ *[I was] a queen in exile* $\xrightarrow{implication}$ *who must bide her time* $\xrightarrow{enables}$ *she could think of some plot, some plan* $\xrightarrow{purpose}$ *return her to her throne*

- one of the guards searched Euphronius he actually put his unclean hands on our wizard's hold person I (Cleopatra Selene II) watched, aghast, trying to ignore the curious motion within the basket an echo of fear that snaked around my heart then the ill-mannered Roman guard approached me and I held my basket out to him hoping he'd reach inside (Counterfactual) hoping that whatever evil spirit lurked there would fly out strike him dead
  one of the guards searched Euphronius $\xrightarrow{efficient}$ *I watched aghast trying to ignore the curious motion within the basket* $\xrightarrow{outcome}$ *the ill-mannered Roman guard approached me* $\xrightarrow{induces}$ *I held my basket out to him* $\xrightarrow{purpose}$ *he'd reach inside* $\xrightarrow{efficient}$ *whatever evil spirit lurked there would fly out* $\xrightarrow{efficient}$ *strike him dead*

6 Conclusion & future direction

For this study, we designed and demonstrated a procedure to rank the likelihood of causality from complex linguistic structures. The process takes lexico-semantic as well as morpho-syntactic information in the expressions into a single form of representation; a collection of which then is extended into a *diffuse prototype*, a composite cognitive categorization model, for a complex multi-modal description of causality. An evolutionary algorithm, with a graph theoretic focus, is developed specifically to obtain the *diffuse prototype* from a limited number of training samples. The output model then can be used to score unseen samples according to a variegated notion of causality. Due to the nature of the model representation and the GA-like procedure, it is adaptable for a wide variety of human definitions of causality. This system in the future needs to be further developed from a ranking procedure to a discrete classification task. It will also be worth to look at further sub-classifications of causality, to see whether a similar procedure can provide a yet more fine-grained recognition of different deep semantic types of this relation.

References

S. Abe, K. Inui and Y. Matsumoto 2008. *Two-phrased event relation acquisition: Couplingthe relation-oriented and argument-oriented approaches. Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, 1−8

L. M. Alcott 1997. *Little women.* Roberts Brothers Publishing, Boston.

M. P. Alvarez 2009. *The four causes of behavior: Aristotle and Skinner. International Journal of Psychology and Psychological Therapy*, 9(1):45−57

Aristotle 300 B.C. φυσικὴ ἀκρόασις (Physics). Translated by R. P. Hardie and R. K. Gaye, MIT Press 1994.

Aristotle before 322 B.C. τὰ μετὰ τὰ φυσικά (Metaphysics), ed. J. Verner. Oxford Classical Texts: Oxford University Press, 1957.

J. E. Baker 1985. *An Analysis of the Effects of Selection in Genetic Algorithms. International Conference on Genetic Algorithms and Their Applications*, 101−111

J. E. Baker 1989. *An Analysis of the Effects of Selection in Genetic Algorithms. Ph.D. Thesis, Vanderbilt University*, Nashville, 1989

N. A. Barricelli 1989. *Numerical testing of evolution theories : Part I Theoretical introduction and basic tests. Acta Biotheoretica*, Issue 16 (1-2):69−98

R. Girju and B. Beamer 2009. *Using a bigram event model to predict causal potential. In Proceedings of Conference on intelligent text processing and computational linguistics 2009*, Mexico City, Feb 26−28

S. Berthard & J. H. Martin 2008. *Learning semantic links from a corpus of parallel temporal and causal relations. Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, 177−180, Association for Computational Linguistics

H. Bodlaender 1988. *Dynamic programming on graphs with bounded treewidth. Proceedings of 15th International Colloquium on Automata, Languages, and Programming*, volume 317 of Lecture Notes in Computer Science, 105118. Springer-Verlag, Berlin Heidelberg

T.C. Bora, L. Lebensztajn, L.D.S. Coelho 2012. *Non-Dominated Sorting Genetic Algorithm Based on Reinforcement Learning to Optimization of Broad-Band Reflector Antennas Satellite. IEEE Transactions in Magnetics*, 48(Number 2 Supplement 4 Part 1):767−770, Piscataway, NY

A. Brindle 1981. *Genetic algorithms for function optimization*, *Technical Report*, 81−82, University of Alberta, Canada

British National Corpus 2007. *University of Oxford Press, Longman Publishing, W & R Chambers Publishing, in conjunction with British Library, University of Oxford, and Lancaster University* url:http://www.natcorp.ox.ac.uk/

B. Chakraborty and P. Chaudhuri 2003. *On The Use of Genetic Algorithm with Elitism in Robust and Nonparametric Multivariate Analysis.* Austrian Journal of Statistics, Volume 32

D.S. Chang, K.S. Choi 2006. *Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities Information Processing and Management*, 42(3):662678

C. Chudasama, S. M. Shah, M. Panchal 2011. *Comparison of parents selection methods of genetic algorithm for TSP. International Conference on Computer Communication and Networks*, published by International Journal of Computer Applications (IJCA)

N. L. Cramer 1985. *A Representation for the adaptive generation of simple sequential programs. Proceedings of International Conference on Genetic Algorithms ans their Applications*, Carnegie-Mellon University

C. Darwin 1859. *on the Origin of species: (by Means of Natural Selection, The Preservation of Favoured Races in the Struggle for Life)* John Murray Publishing

R. Delmonte, A. Bristot, M.A. Piccolino Boniforti, S.Tonelli 2007. *Entailment and anaphora resolution in RTE3. Proceedings of ACL Workshop on Text Entailment and Paraphrasing*, 48−53, Association of Computational Linguistics, Prague

D. B. Fogel 1998. *Evolutionary computation: the Fossil record*, IEEE Press, Piscataway, NJ

R. Girju 2003 *Automatic detection of causal relations for question answering*, *The 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Workshop on Multilingual Summarization and Question Answering : Machine Learning and Beyond

D. Goldberg 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing, Reading MA

D. Goldberg and K. Deb 1993. *A comparative analysis of selection schemes used in genetic algorithms. Foundations of Genetic Algorithms*, 69−93

D. Goldberg, K. Deb, J. Clark 1991. *Genetic algorithms, noise, and sizing of population. Complex System*, 6:333-362

Project Gutenberg, M. Hart 2005. *Gutenberg corpus*, http://www.gutenberg.org/

C. Hashimoto, K. Torisawa, J. Kloetzer, M. SanosIstvan, V. J.H. Oh, Y. Kidawara 2014. *Toward future scenario generation: extracting event causality exploiting semantic relation, context, and association features. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*

J. D. Holland 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, Michigan

K. A. de Jong 1975. *an Analysis of the behavior of a class of genetic adaptive systems*, Ph. D. Dissertation, University of Michigan Press, Ann Arbor, Michigan

C. Khoo, S. Myaeng, R. Oddy 2001. *Using cause-effect relations in text to improve information retrieval precision*, *Information Processing and Management*, 37:119145

P.R. Killeen 2001. *The four causes of behavior*, *Current Directions in Psychological Science*, 10:136−140

P.R. Killeen and M.R. Nash 2003. *The four causes of hypnosis. The International Journal of Clinicaland Experimental Hypnosis*, 51:195−231

S. Kirkpatrick, J. C.D. Gelatt, M. P. Vecchi 1983. *Optimization by simulated annealing. Science*, 220:671−680

S. Mashohor 2005. *Elitist selection schemes for genetic algorithm based printed circuit board inspection system. Evolutionary Computation, The 2005 IEEE Congress*, 2:974−978

P. Menzies 2009. *Counterfactual theories of causation. Stanford Encyclopedia of Philosophy, Fall 2009*, ed: E. N. Zalta, web published: plato.stanford.edu/entries/causation-counterfactual Stanford University Department of Philosophy

J.H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. de Saeger, K. Ohtake 2013. *Why-question answering using intra- andinter-sentential causal relations Proceedings of the 51st Annual Meeting of the Association for Com-putational Linguistics (ACL 2013)*, 51:17331743

M. Paul, C. Girju, C. Li 2009. *Mining the web for reciprocal relationships*, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, Boulder, Colorado

H. Rachlin 1992. *Teleological behaviorism. American Psychologist*, 47:1371-1382, American Psychological Association

K. Radinsky, S. Davidovich, S. Markovitch 2012. *Learning causality for newsevents prediction. Proceedings of International World Wide Web Conference 2012 (WWW 2012)*, 909−918

I. Rechenberg 1973. *Evolutionsstrategic: Optimierung technicher System nach Principen der Biologischen Evolution*, Frommmann-Holzboog Verlag, Stuttgart

M. Riaz and C. R. Girju 2010. *Another look at causality: Discovering scenario-specific contingency relationships with no supervision*, *Proceedings of ICSC*, CERN School of Computing

J. Schmidhuber 1987. *Evolutionary principles in self-referential learning: On Learning now to learn: the meta-meta-meta...*, Doctoral Thesis, Technische Universität Munchen, Germany

A. Turing 1950. *Computing machinery and intelligence Mind: a Quarterly Review of Psychology and Philosophy*, LIX(236):433−460, The Mind Association

F. Yaman, A. E. Yilmaz 2012. *Elitist genetic algorithm performance on the uniform circular antenna array pattern synthesis problem*, *PRZEGLD ELEKTROTECHNICZNY (Electrical Review)*

S. Yang 2007. *Genetic algorithms with elitism-based immigrants for changing optimization problems*, *EvoWorkshops 2007*, LNCS 4448:627−636, Springer Verlag, Berlin.