Ants for Natural Language Processing

Mathieu Lafourcade, F. Guinand

Ants for Natural Language Processing

Mathieu Lafourcade[1] , Frédéric Guinand[2]

[1] LIRMM (CNRS - Université Montpellier 2)

mathieu.lafourcade@lirmm.fr

[2] Laboratoire d'Informatique du Havre

frederic.guinand@univ-lehavre.fr

Abstract

The conceptual vector model aims at representing word meanings by concept activations for Natural Language Processing (NLP) applications like Word Sense Disambiguation (WSD) and Thematic Analysis. Learning capabilities of a system implementing this model is a characteristic needed in the path to natural language comprehension. The analysis process is defined as a vector propagation on a morphosyntactic analysis tree. Relative word sense activations is a basic mean for an efficient WSD. However, such strategy does not take into account interpretation trails, and constraints between word senses. We propose to consider this problem as a dynamic complex system in which moving entities express the interactions between the various elements of the text and stemmed from its analysis. This approach combines ant algorithms and conceptual vectors. Early experiments gave hints that beside being cognitively motivated on some aspects, this approach can perform in some difficult cases very well and is a generalization of the standard propagation.

1 Introduction

In the framework of the Word Sense Disambiguation (WSD) and lexical transfer in Machine Translation (MT), the representation of word meaning is one main issue. The conceptual vector model aims at representing thematic activations for chunks of text, lexical entries, locutions up to whole documents. Roughly speaking, vectors are supposed to encode *ideas* associated to words or expressions. The main applications of the model are thematic text analysis and lexical disambiguation [Lafourcade 2001]. Practically, [Lafourcade *et al.* 2002] presents a system, with automated learning capabilities, based on conceptual vectors and exploiting monolingual dictionaries (available on the web). So far, from French, the system learned 110000 lexical entries corresponding to roughly 430000 vectors (the average meaning number for polysemous word being 5.1). The same experiment is conducted for English. The issue addressed in this paper concerns the analysis process itself.

Understanding a text requires the comparison of the various meanings of any polysemous word with the context of its use. But, the context is precisely defined by the words themselves with all their meanings and their status in the text (verb, noun, adjective...). In a sense, words are the basic compounds of an interaction network which implicit dynamics reveals the pregnancy of each meaning associated to any polysemous word. If we refer to the most commonly shared definition of a *complex system*, it states that it is a *network of interacting entities*, objects, agents, elements, or processes that exhibit a *dynamic*, aggregate *behavior*. The action of an object (possibly) affects subsequent actions of other objects in the network, so that *the action of the whole is more than the simple sum of the actions of its parts.*[1]. The *actions* in our context correspond to the *meanings* of the words constituting the text, and the sum of the actions results in the global meaning of the text, which is, for sure, much more than the simple sum of the meanings of the words. Then, in itself, a text constitutes a complex system. The computational problem is that the meanings are not strictly speaking active elements. In order to ensure the dynamicity of the whole system, an active framework made of ”meaning transporters” must be supplied to the text. These ”transporters” are intended to allow the interactions between text elements. They have to be both light (because of their possible large number) and independent (word meanings are intrinsic values). Moreover, when some meanings stemmed from different words are compatible (*engaged* with *job* for instance), the system has to keep a trace of this fact. This set of constraints has led us to consider ant algorithms. Ants algorithms or variants of them have been classically used for optimization problems; traveling salesman problem (TSP) [Dorigo and Gambardella 1997], graph coloring [Costa and Hertz 1997], routing problems [Caro and Dorigo 1998] [Bruten *et al.* 1996], dynamic load balancing [Bertelle *et al.* 2003], and more recently for computational molecular biology problems; protein identification [Gras *et al.* 2002] or DNA-Sequencing using Sequencing-by-Hybridization method [Bertelle *et al.* 2002a], but they were never used in Natural Language Processing (NLP). Most probably because NLP was neither modeled as an optimization problem, nor explicitly modeled as a dynamic complex system. However, [Hofstadter 1995] with the COPYCAT project, presented an approach where the environment by itself contributed to solution computation and is modified by an agent population where roles and motivations varies. In [Gale 1992], Church and Yarowsky have used Naive-Bayes algorithm for WSD. Some properties of these models seem to be adequate for the task of semantic analysis and WSD, where word senses can be seen as competing for resources. We retain here some aspects that we consider as being crucial: (1) mutual information or semantic proximity is one key factor for lexical activation, (2) the syntactic structure of the text can be used to guide information propagation, (3) conceptual bridges can be dynamically constructed (or deleted) and could lead to *catastrophic events* (in the spirit of [Thom 1972]). These bridges are the instrumental part allowing mutual-information exchange beyond locality horizons. Finally, as pointed by [Hofstadter 1995], biased randomization (which doesn't mean chaos) plays a major role in the model.

In this paper, we first expose the conceptual vectors model and the notions of semantic distance and contextualization. Then, we detail the text analysis process coupled with conceptual vectors, named Standard Propagation (SP), which is used for text classification, thematic analysis and vector learning. After analyzing some drawback

[1] [Langton 1996] *Why do we need artificial life ?* page 305.

of the SP approach, we present the Colored Ant (CA) Algorithm. CA includes SP in its behavior but extends it with interpretation trail creation and prepositional phrases attachment. We show, that the overall process is by essence emergent.

2 Conceptual Vectors

Thematic aspects of textual segments (documents, paragraphs, syntagms, etc.) can be represented by conceptual vectors. Vectors have been used in information retrieval for long [Salton and MacGill 1983] and for meaning representation by the LSI model [Deerwester *et al.* 1990] from latent semantic analysis (LSA) studies in psycholinguistics. In computational linguistics, [Chauché 1990] proposes a formalism for the projection of the linguistic notion of semantic field in a vectorial space, from which this model is inspired. From a set of elementary notions, concepts, it is possible to build vectors (conceptual vectors) and to associate them to lexical items[2]. The hypothesis that considers a set of concepts as a generator to language has been long described in [Rodget 1852] (*thesaurus hypothesis*). Polysemous words combine the different vectors corresponding to the different meanings. This vector approach is based on well known mathematical properties, it is thus possible to undertake well founded formal manipulations attached to reasonable linguistic interpretations. Concepts are defined from a thesaurus (in the prototype applied to French, [Larousse 1992] has been chosen where 873 concepts are identified. This figure is comparable both quantitatively and qualitatively with the thousand defined in [Rodget 1852] for English). To be consistent with the thesaurus hypothesis, we consider that this set constitutes a generator space for the words and their meanings. This space is probably not free (no proper vectorial base) and as such, any word would project its meaning on this space.

2.1 Thematic Projection Principle

Let be $\mathcal{C}$ a finite set of n concepts, a conceptual vector V is a linear combination of elements c_i of $\mathcal{C}$. For a meaning A, a vector $V(A)$ is the description (in extension) of activations of all concepts of $\mathcal{C}$. For example, the different meanings of ‹*quotation*› could be projected on the following concepts (the *CONCEPT*[*intensity*] are ordered by decreasing values): V(‹*quotation*›) = *STOCK EXCHANGE*[*0.7*], *LANGUAGE*[*0.6*], *CLASSIFICATION*[*0.52*], *SYSTEM*[*0.33*], *GROUPING*[*0.32*], *ORGANIZATION*[*0.30*], *RANK*[*0.330*], *ABSTRACT*[*0.25*], ...

In practice, the largest $\mathcal{C}$ is, the finer the meaning descriptions are. In return, the computer manipulation is less easy. It is clear, that for dense vectors[3] the enumeration of the activated concepts is long and difficult to evaluate. We would generally prefer to select the thematically closest terms, i.e., the *neighborhood*. For instance, the closest terms ordered by increasing distance of ‹*quotation*› are: $\mathcal{V}$(‹*quotation*›) = ‹*management*›, ‹*stock*›, ‹*cash*›, ‹*coupon*›, ‹*investment*›, ‹*admission*›, ‹*index*›, ‹*abstract*›, ‹*stock-option*›, ‹*dilution*›, ...

[2] Lexical items are words or expressions which constitute lexical entries. For instance, ‹*car*› or ‹*white ant*› are lexical items. In the following we will use sometimes *word* or *term* to speak about a *lexical item*.

[3] Dense vectors are those which have very few null coordinates. In practice, by construction, all vectors are dense.

2.2 Angular Distance

Let us define $Sim(A, B)$ as one of the *similarity* measures between two vectors A et B, often used in information retrieval. We can express this function as the scalar product of their vector divided by the product of their norm. We suppose here that vector components are positive or null. Then, we define an *angular distance* D_A between two vectors A and B as:

$$\begin{aligned} & D_A(A, B) = \arccos(Sim(A, B)) \\ \text{with} \quad & Sim(A, B) = \cos(\widehat{A, B}) = \frac{A \cdot B}{\|A\| \times \|B\|} \end{aligned} \tag{1}$$

Intuitively, this function constitutes an evaluation of the *thematic proximity* and is the measure of the angle between the two vectors. We would generally consider that, for a distance $D_A(A, B) \leq \frac{\pi}{4}$, (i.e. less than 45 degrees) A and B are thematically close and share many concepts. For $D_A(A, B) \geq \frac{\pi}{4}$, the thematic proximity between A and B would be considered as loose. Around $\frac{\pi}{2}$, they have no relation. D_A is a real distance function. It verifies the properties of reflexivity, symmetry and triangular inequality. We can have, for example, the following angles:

D_A(‹*profit*›, ‹*profit*›)=0°	D_A(‹*profit*›, ‹*product*›)=32°
D_A(‹*profit*›, ‹*benefit*›)=10°	D_A(‹*profit*›, ‹*goods*›)=31°
D_A(‹*profit*›, ‹*finance*›)=19°	D_A(‹*profit*›, ‹*sadness*›)=65°
D_A(‹*profit*›, ‹*market*›)=28°	D_A(‹*profit*›, ‹*joy*›)=39°

The first value has a straightforward interpretation, as ‹*profit*› cannot be closer to anything else than itself. The second and the third are not very surprising since a ‹*benefit*› is quite synonymous of ‹*profit*›, in the ‹*finance*› field. The words ‹*market*›, ‹*product*› and ‹*goods*› are less related, which explains a larger angle between them. The idea behind ‹*sadness*› is not much related to ‹*profit*›, contrary to its antonym ‹*joy*› which is thematically closer (either because of metaphorical meanings of ‹*profit*› or other semantic relations induced by the definitions). The thematic proximity is by no way an ontological distance but a measure of how strongly meanings may relate to each others.

The graphical representations of the vectors of ‹*exchange*› and ‹*profit*› shows that these terms are indeed quite polysemous. Two other terms (‹*cession*› and ‹*benefit*›) seems to be more focused on specific concepts. These vectors are the average of all possible meanings of their respective word in the general Thesaurus. It is possible to measure the level of *fuzziness* of a given vector as a clue of the number of semantic fields the word meaning is related to.

Because of the vagueness related either to polysemy or to lacks of precision (only 873 general concepts), vectors have to be *plunged* into a specialized semantic space. However, one cannot cut loose from the general vectors for two reasons. First, even non-specialized words may turn out to be pivotal in word sense disambiguation of specialized ones. Second, we cannot know beforehand if a given occurrence of a word should be understood in its specialized acception or more a general one.

One would certainly consider that the angle between two vectors can be regarded as a similarity measure, and *of course* the cosine between vectors (or $1 - \cos$ if a metric

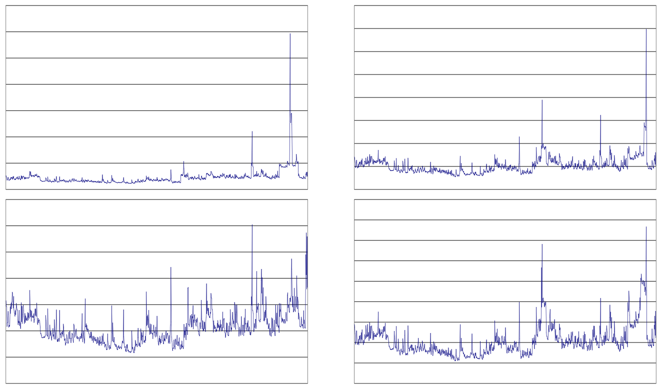

Figure 1: Graphical representation of the vectors of 4 terms: *cession* and *benefit* (top), *exchange* and *profit* (bottom) (rather polysemous)

is required) is also a similarity measure, and that's everyone else uses. This remark is by itself correct, but we would like to stress that using the *arcos* metric lead to a more discriminant function at small angles. Precisely, generally we do consider the limit of the *cos* metric at 0.5, below this value there is no much common information. With *arcos*, we set this limit at $\pi/4$ which corresponds to a value of $\sqrt{2}/2$. What is really important is to be able to discriminate strongly between objects and certainly not in a linear way, as already distant objects are not to be scrutinized contrary to seemingly closed ones.

2.3 Vector Operators

Vector Sum. Let X and Y be two vectors, we define their *normed sum* V as:

$$V = X \oplus Y \quad | \quad v_i = (x_i + y_i)/\|V\| \tag{2}$$

This operator is idempotent and we have $X \oplus X = X$. The null vector $\vec{\mathbf{0}}$ is by definition the neutral element of the vector sum. Thus we write down that $\vec{\mathbf{0}} \oplus \vec{\mathbf{0}} = \vec{\mathbf{0}}$.

Normed Term to Term Product. Let X and Y be two vectors, we define V as *their normed term to term product*:

$$V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i} \tag{3}$$

This operator is idempotent and $\vec{\mathbf{0}}$ is absorbent. We have: $V = X \otimes X = X$ and $V = X \otimes \vec{\mathbf{0}} = \vec{\mathbf{0}}$.

Contextualization. When two terms are in presence of each other, some of the meanings of each of them are thus selected by the presence of the other, acting as a context. This phenomenon is called *contextualization*. It consists in emphasizing common features of every meaning. Let X and Y be two vectors, we define $\Gamma(X,Y)$ as the contextualization of X by Y as:

$$\Gamma(X,Y) = X \oplus (X \otimes Y) \tag{4}$$

These functions are not symmetrical. The operator Γ is idempotent ($\Gamma(X, X) = X$) and the null vector is the neutral element. ($\Gamma(X, \vec{\mathbf{0}}) = X \oplus \vec{\mathbf{0}} = X$). We will notice, without demonstration, that we have thus the following properties of *closeness* and of *distance*):

$$\begin{aligned} & D_A(\Gamma(X, Y), \Gamma(Y, X)) \\ \leq \{ & D_A(X, \Gamma(Y, X)), D_A(\Gamma(X, Y), Y)\} \leq D_A(X, Y) \end{aligned} \tag{5}$$

The function $\Gamma(X, Y)$ brings the vector X closer to Y proportionally to their intersection. The contextualization is a low-cost way of amplifying properties that are salient in a given context. For a polysemous word vector, if the context vector is relevant, one of the possible meanings is *activated* through contextualization. For example, *bank* by itself is ambiguous and its vector is pointing somewhere between those of *river bank* and *money institution*. If the vector of *bank* is contextualized by *river*, then concepts related to finance would considerably be dimmed.

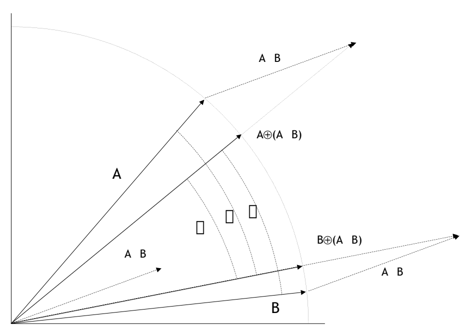

Figure 2: Geometric representation (in 2D) of the contextualization function. The α angle represents the distance between A and B contextualized by each other.

2.4 From Text to Vectors by Standard Propagation

How do we build a conceptual vector from a given text? From this text, we first compute a morphosyntactic analysis tree. This is a derivation tree from where leaves (roughly) reconstitute the original sentence. A leaf refers to a word on which are associated one or several definitions (as found in dictionaries) and a conceptual vector. For simplicity, we only consider contents words, that is nouns, verbs, adjectives and adverbs. After filtering according to agreement on morphosyntactic attributes, is attached to the leaf a *uncontextualized global* conceptual vector computed from the vectors of its k definitions. The most straightforward way (not the best) to do so is to compute the average vector: $V(w) = V(w.1) \oplus \cdots \oplus V(w.k)$. If the word is unknown (i.e. it is not in the dictionary), the null vector is taken instead.

Vectors are then propagated upward. Consider a tree node N with p dependents $N_i(1 \leq ip)$. The newly computed vector of N is the weighted sum of all the vector of N_i: $V(N) = \alpha_i N_1 \oplus \cdots \oplus \alpha_p N_p$. The weights α depend of the syntactic functions of the node. For instance a governor would be given a higher weight ($\alpha = 2$) than a regular node ($\alpha = 1$) so that, for example, the vectors computed for ***a boat sail*** and for ***a sail boat*** would not be identical. Once the vector of the tree root is determined a downward propagation is performed. A node vector is contextualized by its parent: $V'(N_i) = V(N_i) \oplus \Gamma(N_i, N)$. This descent is done recursively until reaching a leaf. At the leaf level an implicit WSD process is undertaken as the new *contextualized global* vector is then a weighted sum of the vector of the definitions where weights are non-linearly related to the amount of mutual information between the context (node N) and a given meaning:

$$\begin{aligned} V'(w) &= \beta_i V(w.1) \oplus \cdots \oplus \beta_i V(w.k) \\ \text{with} \qquad \beta_i &= \cot(D_A(V(N), V(w.i)) \end{aligned} \tag{6}$$

If the vector context $V(N)$ is very close to $w.i$, then the global vector $V(w)$ for the word w is almost equal to $V(w.i)$ (we recall that *cot* refers to the *cotangent* function, with $\cot(0) = +\infty \quad$ and $\quad \cot(\pi/2) = 0$).

The processes of upward and downward propagation are iterated until either a maximum number of cycles is reached or when the root vector stabilizes (it is proved that in all generality there is no convergence as sometimes oscillations happen with strongly ambiguous sentences).

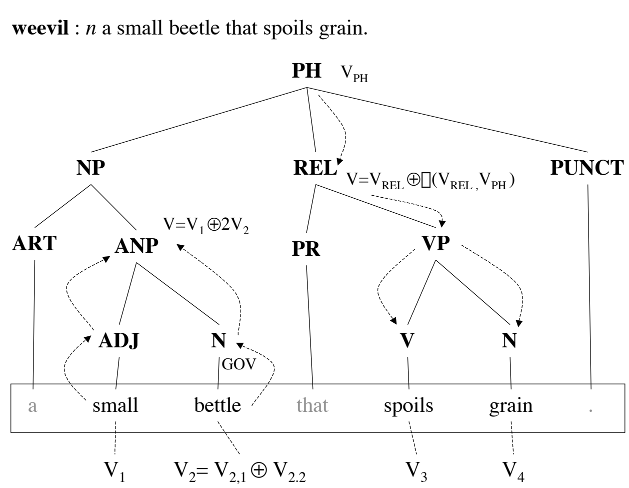

Figure 3: Simplified graphical representation of conceptual vector upward and downward propagation.

One major drawback of this analysis model is that the various interpretations are merged together, and constraints between selected word senses are not structurally represented.

2.5 Conceptual Vector Learning

Before processing texts, a lexicon associating terms and vectors should be constructed. Beside full manual indexing, which is difficult and time consuming, a supervised learning can be devised.

The learning process is an *ever-going* task that consists in picking randomly a word to be learned (or revised). The vector of each definition of this word is then computed as described above. At the beginning the dictionary is empty, and a bootstrapping is done by manually indexing a small set of common words

3 Colored Ants for Word Sense Disambiguation

Ant algorithms are a class of meta-heuristics based on a population of individuals exhibiting a cooperative behavior[Langton 1987]. Ants continuously forage their territories to find food[Gordon 1995] visiting paths, creating bridges, constructing nests, etc. A fundamental principle in the emergence of coordinated system-level behavior from the local interactions of ants is stigmergy. This concept was introduced by Grassé in the 1950s from the interpretation of the behavior of social insects [Grassé 1959]. The idea of stigmergy is that a collaborative task (clustering, nest building, food search...) is implicitly coordinated through the elements (signs and/or modifications) resulting from individuals activities. For instance, ants perform indirect communications using chemical signals called *pheromones*. The larger the quantity of pheromones on a path, the larger the number of ants visiting this path. Thus, signs left on paths by some ants influence choices of next ants. This characteristic was successfully exploited for processing various combinatorial optimization problems like TSP or routing in networks [Dorigo and Gambardella 1997, Di Caro *et al.* 1998]. However, two general forms of stigmergy are identified. The first is sign-based stigmergy. The elements left by ants in the environment don't directly contribute to the achievement of the collaborative task. Pheromones fall into this category. The second form is sematectonic stigmergy. It generally involves a change in the physical characteristics of the environment. Elements of sematectonic stigmergy may be environmental modifications that directly concern the collaborative task. The application of clustering described in [Lumer and Faieta 1994] is based on that kind of stigmergy.

Our method for WSD relies on both kind of stigmergy. Sign-based stigmergy plays a role in ant behaviors. Sematectonic stigmergy is used for modifying nodes characteristics and for creating new paths between vertices. In the sequel, these new paths will be called bridges.

3.1 Motivation for Colored Ants

The "binary bridge" is an experiment developed by [Pasteels *et al.*]. As reported in [Dorigo *et al.* 1999] *in this experiment, a food source is separated from the nest by* *a bridge with two equally long branches A and B.*Initially, both paths are visited and after some iterations, one path is selected by the ants, whereas the second, although as good as the first one, is deserted. This experiment interests us for two reasons. It first shows that ants have the ability of organizing themselves in order to determine a global solution from local interactions, thus, it is likely to obtain an emergent solution for a problem submitted to an ant-based method. This point is crucial for our problem, since we expect the emergence of a meaning for the analyzed text. But, the experiment also shows the inability of such method, in its classical formulation, to provide a set of simultaneous and distinct solutions instead of only one at a time. As these methods are based on the reinforcement of the current best solution, they are not directly suitable for our situation. Indeed, if several meanings are possible for a text, all these meanings should emerge. In this work we present an ant-based method implementing several colonies competing for promoting their meaning and collaborating for the building of global meanings. These colonies are distinguished by colors: one color is associated to each sense of each term. A similar approach was already successfully implemented for performing dynamic load balancing in the context of simulations [Bertelle *et al.* 2003]. These competing colonies provide their own solution constrained by the senses of the terms forming the text. Usually the emergence of a global solution results from local interactions, and in most cases local interactions are limited to geographic proximity. In our case, local interactions should also concern semantic proximity of words that may be very distant in the text and in the morphosyntactic tree. For that reason, in our model ants are allowed to modify their environment by building bridges between "friends words" as it will be described in the sequel.[4] Text analysis results in computing a global conceptual vector at each level of the analysis tree. These vectors represent the *ideas* expressed at various text granularity: terms, groups, sentences, paragraphs... At the root level, we want to get the global concept activations of the entire text. These goals are extended to get some explicit interpretation representation, in the form of transversal trails between selected word senses. In the French sentence *L'avocat est véreux.* (see figure 4) we have only two reasonable interpretations (out of four). Generally, not all combinations of word senses are possible and the Standard Propagation Analysis only accounts for conceptual activation not *interpretation trails*.

3.2 Environment

As before, the underlying structure of the environment is the morphosyntactic analysis tree of the text to be analyzed. Each content word is a node. This node has as many children in the tree as senses. To each child associated to a sense corresponds a unique color: the conceptual vector of the sense. A child is also a *nest* and all children of a node associated to a content word are *competing nests*. In figure 4, both nodes are directly linked to two nests. An ant can walk through graph edges and, under some circumstances, can build new ones (called bridges). Each node contains the following attributes beside the morphosyntactic information computed by the analyzer: (1) a resource level R, and (2) a conceptual vector V. Each edge contains (1) a pheromone

[4] friend senses ? Senses which conceptual vectors are close in term of angular distance will be called friends in the sequel

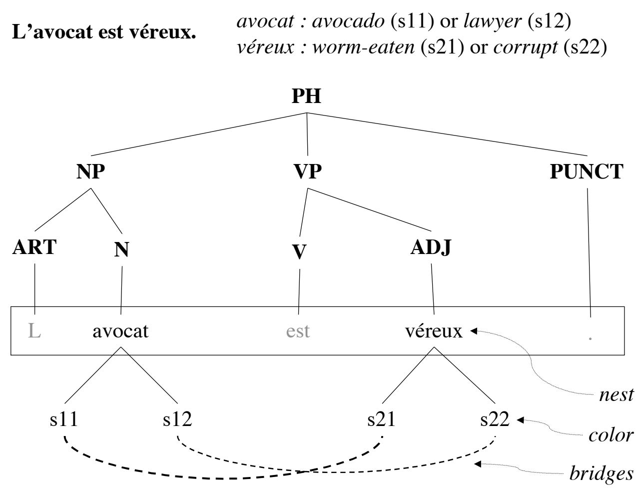

Figure 4: For the french sentence *L'avocat est véreux.* There are two interpretations (*The avocado is worm-eaten* or *The lawyer is corrupt*. An interpretation trails is a strongly excited bridge (or bridge sequence) between activated word senses.

level. The main purpose of pheromone is to evaluate how popular a given edge is. The environment by itself is evolving in various aspects:

1. the conceptual vector of a node is slightly modified each time a new ant arrives. Only vectors of nests are invariant (they cannot be modified). A nest node is initialized with the conceptual vector of its word sense, other nodes with the null vector.

2. resources tend to be redistributed toward and between nests which *reinvest* them in ants production. Nodes have an initial amount of resources of 1.

3. the pheromone level of edges are modified by ant moves. There is a factor decay δ (the evaporation factor) which ensures that with time pheromone level tends to decrease toward zero if no ant are passing through. Only bridges (edges created by ants) would disappear if their pheromone level reaches zero.

The environment has an impact on an ant and in return ants continuously modify the environment. The results of a simulation run are decoded thanks to the pheromone level of bridges and the resource level of nests.

3.3 Nests, Ant Life and Death

A nest (word sense) has some resources which are used for producing new ants. The level of resources denoted $R \in [-\infty, +\infty]$ may be negative. However, a nest with a

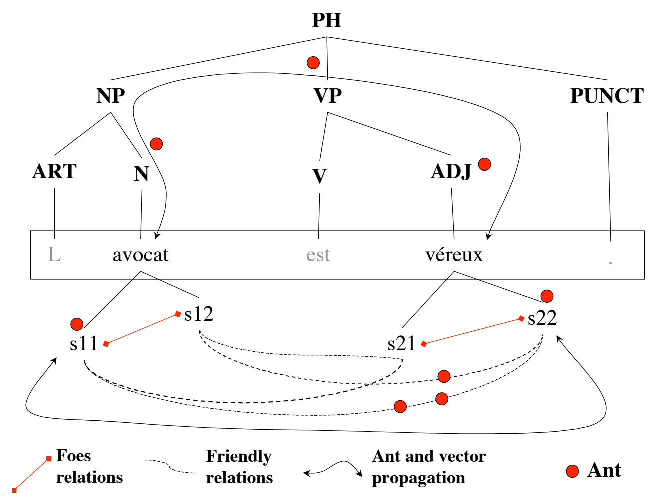

Figure 5: Propagation schema. Ants by their foraging activities establish Friend and Foes relations between nests and are redistributing resources and propagating conceptual vectors among the nodes of the tree. The tree is becoming a graph through the creation of new relations (bridges). Mostly activated relations lead to interpretation trails.

negative level of resources may still produce new ants. At each cycle, among the set of nests having the same parent node (content word), only one is allowed to produced a new ant. The color of this ant is the one of the selected nest. In all generality, a content word has n children (nests), and the nest chosen for producing the next ant is probalistically selected according to the level of resources.

There is a cost ϵ for producing an ant, which is deducted from the nest resources. Resource levels of nests are modified by ants.

The probability of producing an ant, is related to a sigmoid function (see figure 6) applied to the resource level of the nest. The definition of this function ensures that a nest has always the possibility to produce a new ant although the chances are low when the node is inhibited (resources below zero). A nest can still borrow resources and thus a word meaning has still a chance to express itself even if the environment is very unfriendly.

The ant cost can be related to the ant life span λ which is the number of cycles the ant can forage before dying. Below, we discuss the effect of setting a high or low λ on the overall process. When an ant dies, it gives back all the resources it carries plus its cost, to the currently visited node. For instance, if the ant carrying 0.5 dies on node N, the new level of resources of this node is increased by $\epsilon + 0.5$. This approach leads to a very important property of the system, that the total level of resources is constant. The resources can be unevenly distributed among nodes and ants and this distribution changes over time, sometimes leading to some stabilization and sometimes leading to periodic configurations. This is this *transfer of resources* that reflects the lexical

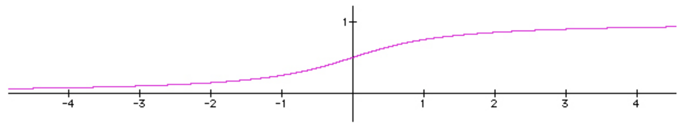

Figure 6: Sigmoid function: $\text{Sig}(x) = \frac{1}{\pi}\arctan(x) + 0.5$. Some values are: $\text{Sig}(0) = 1/2, \text{Sig}(1) = 0.75, \text{Sig}(2) = 0.852, \text{Sig}(-1) = 0.25, \text{Sig}(-2) = 0.147$.

selection, through word senses activation and inhibition.

The ant population (precisely the color distribution) is then evolving in a different way of classical approaches ([Dorigo and Gambardella 1997]) where ants are all similar and their number fixed in advance. However, at any time (greater than λ), the environment contains at most λ ants that have been produced by the nests of a given content word. It means that the global ant population size depends on the number of content words of the text to be analyzed, but not on the number of word meanings. To our views, this is a very strong point that reflects the fact some meanings will express more than others, and that, for a very polysemic word, the ant struggle will be intense. A monosemic word will often serve as a pivot to other meanings. Moreover, this characteristic allows us to evaluate the computing requirements needed for computing the analysis of a given text since the number of ants depends only on the number of words.

3.4 Ant Population

An ant has only one motivation: foraging and bringing back resources to its nest. To this purpose, an ant has two kinds of behavior (called modes), (1) searching and foraging and (2) returning resources back to the nest. An ant a has a resource storage capacity $\text{R}(a) \in [0, 1]$. At each cycle, the ant will decide between both modes as a linear function of its storage. For example, if the $\text{R}(a) = 0.75$, there is a 75% chance that this ant a is in *bringing back* mode.

Each time an ant visits a (non-nest) node, it modifies the node color by adding a small amount of its own color. This modification of the environment is one factor of the sematectonic stigmergy previously mentioned and is the means for an ant to find its way back home. The new value of the color is computed as follows: $C(N) = C(N) + \alpha C(a)$ with $0 < \alpha < 1$. In our application, colors are conceptual vectors and the "+" operation is a normalized vector addition ($V(N) = V(N) \oplus \alpha V(a)$). We found heuristically, that $\alpha = 1/\lambda$ constitutes a good trade-off between a static and a versatile environment.

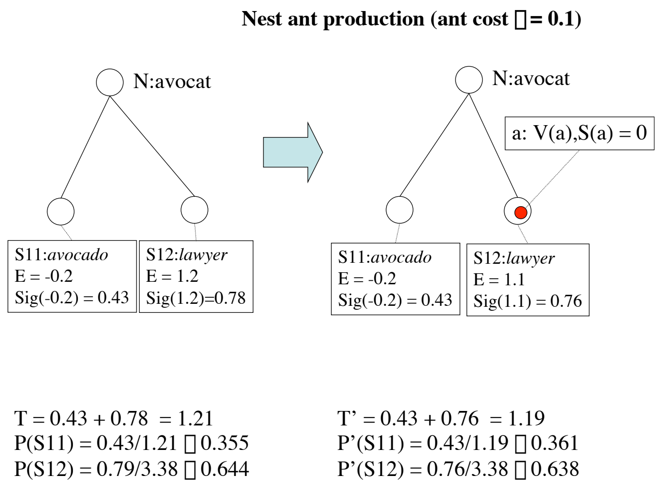

Figure 7: Example of nest ant production. Nest *S12* has been selected to produce a new ant, thus decreasing its total amount of resources by ϵ. For the next cycle, the respective probability of ant production between *s11* and *S12* have been adjusted.

3.5 Searching Behavior

Given a node N_i. N_j is a neighbor of N_i if and only if there exists an edge E_{ij} linking both nodes. A node N_i is characterized by a resource level noted as R(N_i). An edge E_{ij} is characterized by a pheromone level noted as Ph(E_{ij}). A searching ant will move according to the resource level of each neighboring node (its own nest excepted) and to the level of pheromones of the outgoing edges. More precisely an attraction value is computed for each neighbor. This value is proportional to the resource level and inversely proportional to the pheromone level:

$$\text{attract}_S(N_x) = \frac{\max\left(\text{R}(N_x), \eta\right)}{\text{Ph}(E_{ix}) + 1} \tag{7}$$

Where η is a tiny constant avoiding null values for attraction. The motivation for considering an attraction value proportional to the inverse of the pheromone level is to encourage ants to move to non visited parts of the graph. If an ant is at node N_i with p neighbors $N_k(k = 1 \cdots p)$, the probability $P_S(N_x)$ for this ant to choose node N_x in *searching* mode is:

$$P_S(N_x) = \frac{\text{attract}_S(N_x)}{\sum_{1 \leq j \leq p} \text{attract}_S(N_j)} \tag{8}$$

Then, if all neighbors of a node N_i have the same level of resources (including zero), then the probability for an ant visiting N_i to move to a neighbor N_x depends only on the pheromone level of the edge E_{ix}.

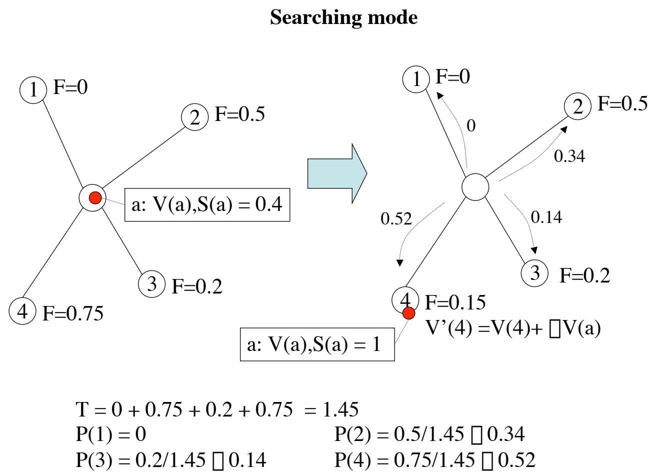

Figure 8: Example of node selection for a searching ant. The node 4 have been (randomly) chosen inducing a color propagation of the ant color.

An ant is attracted by node with a large supply of resources, and will take as much as it can hold (possibly all node resources, see figure 8). A depleted node does not attract searching ants. The principle here, is a simple greedy algorithm.

3.6 Bringing Back Behavior

When an ant has found enough resources, it tends to bring them back to its nest. The ant will try to find its way back thanks to the color trail left back during previous moves. This trail could have been reinforced by ants of the same color, or inversely blurred by ants of other colors.

An ant a returning back and visiting N_i will move according to the color similarity of each neighboring node N_x with its own color and according to the level of pheromones of the outgoing edges. More precisely an attraction value is computed for each neighbor. This value is proportional to the similarity of colors and to the pheromone level:

$$\text{attract}_R(N_x) = \max(\text{sim}(\text{colorOf}(N_x), \text{colorOf}(a)), \eta) \times (\text{Ph}(E_{ix}) + 1) \quad (9)$$

Where η is a tiny constant avoiding null values for attraction.

If an ant is at node N_i with p neighbors $N_k(k = 1 \cdots p)$, the probability $P_B(N_x)$ for this ant to choose node N_x in *returning* mode is:

$$P_R(N_x) = \frac{\text{attract}_R(N_x)}{\sum_{1 \leq j \leq p} \text{attract}_R(N_j)} \tag{10}$$

In our case where colors are represented through conceptual vectors, the similarity function is called the mutual information denoted as mi and defined as follows:

$$\text{mi}(N_x, a) = 1 - \frac{2 \times D_A(\text{vectorOf}(N_x), \text{vectorOf}(a))}{\pi} \tag{11}$$

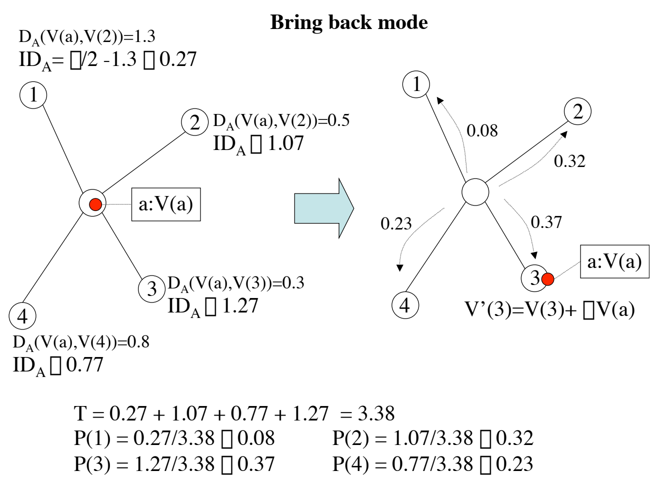

Figure 9: Example of node selection for a bringing ant. The node 3 have been (randomly) chosen inducing a vector propagation of the ant vector.

All considered nodes are those connected by edges in the graph. Thus, the syntactic relations, projected into geometric neighborhood on the tree, dictate constraints on the ant possible moves. However, when an ant is at a friendly nest, it can create a shortcut (called a *bridge*) directly to its home nest. That way, the graph is modified and this new arc can be used by other ants. These arcs are evanescent and might disappear when the pheromone level becomes null.

From an ant point of view, there are two kinds of nests: friend and foe. Foe nests correspond to alternative word senses and ants stemmed from these nests are competing for resources. Friendly nests are all nests of other words. Friends can fool ants by inciting them to give resource. Foe nests instead are eligible as resource sources, that is to say an ant can steal resources from an enemy nest as soon as the resource level of the latter is positive.

3.7 Stigmergy

The stigmergy principle is expressed through color propagation and pheromone deposit, both induced by the ant wandering, and through bridge creation. If ants of a given color tend to be largely present in some part of the tree, then this color will be strongly present in nodes of that region. Ants of other colors will mitigate the colors of such nodes.

During its way back home, an ant may arrive to a friendly nest N. In such a case, the ant gives some part of the resources it carries and may build a bridge from this nest to its own nest. Created bridges constitute one manifestation of sematectonic stigmergy. These bridges play a crucial part in interpretation trails detection since these trails are materialized by a couple of nests linked by a bridge.

The part of the resources left in the friendly nest is proportional to $\mathrm{mi}(N, a)$. For instance, if a is carrying 0.5 and reaches mistakenly a node with $\mathrm{mi}(N, a) = 0.6$, then the ant will give $0.5 * 0.6 = 0.3$ and will have 0.2 resource left. We notice that, when the friendly nest is the home, all resource is given since $\mathrm{mi}(\mathrm{home}, a) = 1$.

Bridges creation can induce *catastrophic events* (in the sense given by [Thom 1972]). Indeed, once created, a bridge allows some ants to reach parts of the tree unreachable otherwise. Note that the creation of a new bridge may change dramatically ant circulation in a very short time, and may ruin all structures established so far.

4 Discussion

4.1 WSD and Interpretation Trails

Correct word senses are selected among the most activated ones. Furthermore, a trail (sequence of nests linked by bridges) between word senses should be present. In rare cases, we may have conflicting results here but this is, most of the time, very significant on the semantic structure, that is, several interpretations are possible for the sentence. Very ambiguous or even humorous sentences, lead to such conflicts and can be detected.

With such a sentence *The old musician donated his organ to the hospital*, we have a strong ambiguity with the word *organ*. Quickly (after 100 cycles in our experiments), the musical instrument interpretation is supported by *musician*, but the presence of *hospital* gives credit to the body part. Both senses of *organ* are activated, but there is not a continuous trail of interpretation for the whole sentence. But a third actor comes into play with *donate*. It doesn't interact much with other words but only slightly with *organ:body part*. After some time (around 400 cycles) a bridge between *donate* and *organ:body part* is able to maintain itself, forcing the interpretation *organ:music* to slowly steps back.

4.2 Prepositional Phrase Attachment

Without adding much to the model, our approach can solve some prepositional phrase (PP) attachment. Consider the very classical example: *He saw the girl in the park*

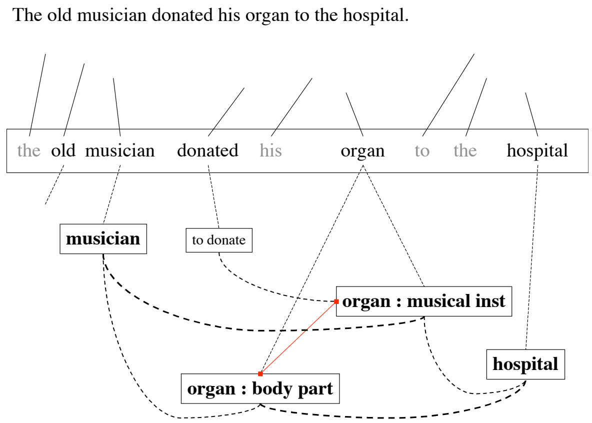

Figure 10: Example of discontinuous trail partially solved by an external weak interaction (originating from the *donate* nest).

with a telescope. First, a strong trail between *saw:see* will be created will *telescope* inducing a strong activation of this sense compared to *saw:saw* (see figure 11).

The only requirement is to enumerate all syntactically acceptable attachments for a PP. Ants and vector propagation will *semantically* choose those which maximize mutual-information sharing. In the sentence *They hit the man with a cane*, the syntagm *with a cane* will be preferably attached to *hit*. Note here, that we are not pretending to actually *solve* any ambiguity, but instead the system computes preferences. These results emerge by the mutual interaction of ants over the environment.

4.3 Results

The evaluation of our model in terms of linguistic analysis is by itself challenging. Manually examining the ant population and node activation on a given text is time consuming. To have a larger scale assessment of our system, we prefer to evaluate it through a Word Sense Disambiguation task (WSD).

A set of 100 small texts have been constituted and each term (noun, adjective, adverb and verb) has been manually tagged. A tag is a term that names one particular meaning. For example, the term *bank* could be annotated as *bank/river*, *bank/money institution* or *bank/building* assuming we restrict ourselves to three meanings. In the conceptual vector database, each word meaning is associated to at least one tag (in the spirit of [Jalabert and Lafourcade 2002]). Using tag is generally much easier than sense number especially for human annotators.

The basic procedure is quite straightforward. The unannotated text is submitted to the system which annotates each term with the guessed meaning. This output is compared to the human annotated text. For a given term, the annotation available to the human annotator are those provided by the conceptual vector lexicon (i.e. for

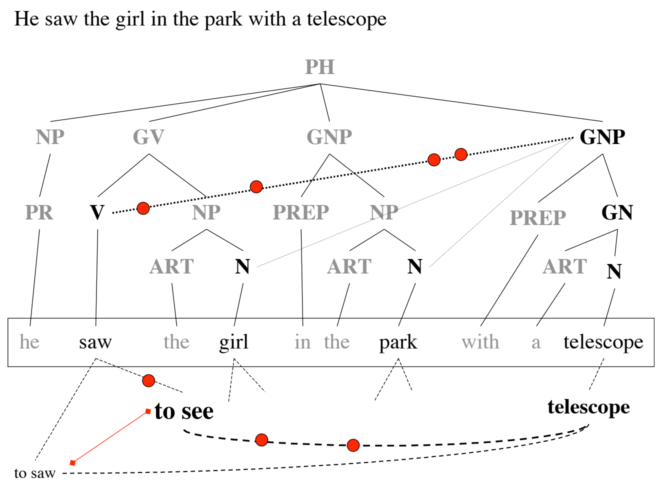

Figure 11: Example of induced PP attachment. Syntactically possible attachments of *with a telescope* are enumerated. The ant population dynamically chooses the shortest path according to conceptual vector mutual information. The strongest trail links *with a telescope* to *saw*. The process is entirely emergent.

bank the human annotator should choose between *bank/river*, *bank/money institution* or *bank/building*). It is allowed for the human annotator to add several tags, in case several meanings are equally acceptable. For instance, we can have *The frigate/*{*modern ship/ancient ship*} *sunk in the harbor.*, indicating that both meanings are acceptable, but excluding *frigate/bird*. Thereafter, we call *gold standard* the annotated text. We should note that only annotated words of the gold standard are target words and used for the scoring.

When the system annotates a text, it tags the term with all meanings which activation level is above 0. That is to say that inhibited meanings are ignored. The system associates to each tag the activation level in percent. Suppose, we have in the sentence *The frigate sunk in the harbor.* an activation level of respectively 1.5 , 1 and -0.2 for respectively *frigate/modern ship*, *frigate/ancient ship* and *frigate/bird*. Then, the output produced by the system is:

> *The frigate/*{*modern ship:0.6/ancient ship:0.4*}.

Precisely, we have conducted two experiments with two different ranking methods.

> A *Fine Grained* approach, for which only the first best meaning proposed by the system is chosen. If the meaning is one of the gold standard tag, the answer is considered as valid and the system scores 1. Otherwise, it is considered as erroneous and the system scores 0.
>
> A *Coarse Grained* approach, more lenient, gives room to closely related meanings. If the first meaning is the good one, then the system scores 1.

Otherwise, the system scores the percent value of a good answer if present. For example, say the system mixed up completely and produced:

The frigate/$\{$*bird:0.8/ancient ship:0.2*$\}$.

the system still gets a 0.2 score.

Scoring scheme	**All terms**	**Nouns**	**Adjectives**	**Verbs**	**Adverbs**
Fine Grain Scoring	0.68	0.76	0.78	0.61	0.85
Coarse Grain Scoring	0.85	0.88	0.9	0.72	0.94

These results compare quite favorably to other WSD systems as evaluated in SENSEVAL campaign [Senseval 2000]. However, our experiment is applied to French which figures are not available in Senseval-2 [Senseval 2 2001].

As expected, verbs are the most challenging as they are highly polysemous with quite often subtle meanings. Adverbs are on the contrary quite easy to figure when polysemous.

We have manually analyzed failure cases. Typically, in most cases the information that allows a proper meaning selection are not of thematic value. Other criteria are more prevalent. Lexical functions, like hyperonymy (is-a) or meronymy (part-of) quite often play a major role. Also, meaning frequency distribution can be relevant. Very frequent meanings can have a boost compared to rare ones (for example with a proportional distribution of initial resources). Only if the context is strong, then could rare meanings emerge.

All those criteria were not modeled and included in our experiments. However, the global architecture we propose is suitable to be extended to ants of other *caste*. In the prototype we have developed so far, only one caste of ants exists, dealing with thematic information under the form of conceptual vectors. Some early assessments seem to show that only with a semantic network restricted *part-of* and *is-a* relations, a 10% gain could be expected (roughly a of gain 12% and a lost of 2%).

5 Conclusion

The conceptual vector model constitutes a numerical approach to lexical semantic representation that is applied to WSD. Contrary to traditional vector models, components refer to ideas or concepts and not to lexical items. More specifically, the methodology includes an autonomous learning of the vectors by the system. Learning is done through the analysis of various lexical information with a strong focus on human usage dictionary definitions.

We stressed on the question of the analysis process. After a first evaluation of a vector propagation over a morphosyntactic analysis tree, it appears that such a simple strategy was falling short in many cases. Mainly, there is a need to explicitly represent connections between selected word senses, thus leading to the creation of interpretation trails. These trails are based on bridges which *solidity* is directly related on their utility for ants.

The ant approach is computationally intensive, but easily parallelizable. The main point is to maintain a (real or simulated) asynchronous parallelism. This parallelism induces partly a biased randomness, which is mandatory for building improbable bridges that may turn out to be very successful. From another spotlight, randomness and bridges may help crossing potential barriers that Standard Propagation cannot cope with. Globally, the model leads to an any-time process that is robust and adaptive. It is possible to add a new sentence next an old one, and watch the previous equilibrium shifting to a new interpretation.

We strongly believe that our approach, in its principles, is potentially very fruitful for semantic analysis. The simplified model presented here only retains the most profound aspects and some extensions have to be done in the way to a very efficient WSD. For instance, all ants are equally competent, their differences being only the color and the position of their nest. It seems clear, that several types (or castes to refer to [Bertelle *et al.* 2002a]) of ants with different linguistic competencies are desirable. Potentially, other phenomena, like anaphoric relations, could then be concurrently tackled by specialized ants.

References

[Langton 1987] C. G. Langton Artificial Life. Addison Wesley.

[Gordon 1995] D. M. Gordon The expandable network of ant exploration *Animal Behaviour* 50:995-1007, 1995.

[Bertelle *et al.* 2002a] C. Bertelle, A. Dutot, F. Guinand, and D. Olivier. DIMANTS: a Distributed Multi-Castes Ant System. In proceedings of Bixmas (Workshop of AAMAS 2002), pages 1-6. Bologna (Italy), July 12-14, 2002.

[Bertelle *et al.* 2003] C. Bertelle, A. Dutot, F. Guinand, D. Olivier Dynamic Placement Using Ants for Object-Based Simulations. LNCS 2888:1263-1274, R. Meersman et al. Editors. Proceedings of CoopIS/DOA/ODBASE 2003, Catania (Sicily), October 2003. Springer-Verlag (Berlin - Heidelberg 2003).

[Bruten *et al.* 1996] J. Bruten R. Schoonderwoerd, O. Holland and L. Rothkrantz. Ant-based load balancing in telecommunications networks. *Adaptative behavior*, 5:169–207, 1996.

[Caro and Dorigo 1998] G. Di Caro and M. Dorigo. AntNet: distributed stigmergetic control for communications networks. *Journal of Artificial Intelligence Research*, 9:317–365, 1998.

[Costa and Hertz 1997] D. Costa and A. Hertz. Ants can color graphs. *Journal of Operation Research Society*, 48:105–128, 1997.

[Chauché 1990] Chauché J., “Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance”, *TA Information*, vol. 31, n° 1, p. 17-24.

[Deerwester *et al.* 1990] Deerwester S., S. Dumais , T. Landauer, G., R. Furnas, Harshman, "Indexing by latent semantic analysis", *Journal of the American Society of Information Science*, 416(6), p. 391-407, 1990.

[Di Caro *et al.* 1998] G. Di Caro and M. Dorigo, "AntNet: Distributed Stigmergetic Control for Communications Networks", *Journal of Artificial Intelligence Research*, vol. 9, p. 317-365, 1998. citeseer.ist.psu.edu/article/dicaro98antnet.html

[Dorigo and Gambardella 1997] Dorigo M., and L. Gambardella, "Ant colony system : A cooperative learning approach to the travelling salesman problem.", *IEEE Transactions on Evolutionary Computation*, 1(1), p. 114-128, 1997.

[Dorigo *et al.* 1999] M. Dorigo E. Bonabeau and G. Theraulaz. *Swarm Intelligence: from natural to artificial systems*. Oxford University Press Inc., 1999. ISBN: 0-19-513158-4.

[Gale 1992] Gale W., K. W. Church, and D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus.", Computers and the Humanities, 26:415-439, 1992.

[Gras *et al.* 2002] R. Gras P. Hernandez and R. D. Appel. Ant colony optimization metaheuristic applied to automated protein identification from tandem mass spectrometry data. In *Proceedings of NETTAB 2002 (Network Tools and Applications in Biology*, Bologna, Italy, July 12th - 14th 2002. http://www.nettab.org.

[Grassé 1959] P.-P. Grassé. La reconstruction du nid et les coordinations inter-individuelles chez Belicositermes natalensis et Cubitermes S.P. La théorie de la Stigmergie : essai d'interprétation du comportement des termites constructeurs. In Insectes Sociaux, vol. 6, pp. 41-80. 1959.

[Hofstadter 1995] Hofstadter, D. R., "Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought (together with the Fluid Analogies Research Group)", NY: Basic Books, 1995.

[Lafourcade *et al.* 2002] Lafourcade M., V. Prince and D. Schwab, "Vecteurs conceptuels et structuration émergente de terminologie", *TAL*, vol 43 - n° 1, p. 43-72, 2002.

[Lafourcade 2001] Lafourcade M., "Lexical sorting and lexical transfer by conceptual vectors", *First International Workshop on MultiMedia Annotation* (MMA'2001), Tokyo, 6 p, January 2001.

[Larousse 1992] Larousse, *Thésaurus Larousse - des idées aux mots - des mots aux idées*. Larousse, 1992, ISBN 1264-4242.

[Langton 1996] C. G. Langton, editor. *Artificial Life: an overview*. MIT, 2nd edition, 1996. ISBN:0-262-12189-1.

[Lumer and Faieta 1994] E. Lumer and B. Faieta, Diversity and Adaptation in Populations of Clustering Ants. Proceedings of the Conference on Simulation of Adaptive Behaviour: from animals to animats 3, pp. 501-508, MIT Press Cambridge, 1994.

[Pasteels *et al.*] J.M. Pasteels J.L. Deneubourg, S. Goss, D. Fresneau, and J.P. Lachaud. Self-organization mechanisms in ant societies (ii): learning in foraging and division of labour. *Experientia Supplementa*, 54:177–196, 1987.

[Rodget 1852] *Thesaurus of English Words and Phrases.* Longman, London, 1852.

[Salton and MacGill 1983] Salton G., MacGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[Thom 1972] Thom R., *Stabilité structurelle et Morphogénèse*, InterEditions, Paris, 1972.

[Yarowsky 1992] Yarowsky D., “Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora”, *COLING’92*, Nantes, p. 454-460, 1992.

[Ide and Véronis2002] Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics 24(1): 1-40 (1998)

[Senseval 2000] http://www.itri.brighton.ac.uk/events/senseval/

[Senseval 2 2001] http://www.sle.sharp.co.uk/senseval2/

[Jalabert and Lafourcade 2002] From sense naming to vocabulary augmentation in Papillon. In proc. of PAPILLON-2003, Sapporo, Japan, July 2002, 12 p.