ATINER CONFERENCE PAPER SERIES No: COM2013-0595

Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
COM2013-0595

Topological Proof of the Computability of the Algorithm based on the Morphosyntactic Distance

Pedro G. Guillén
Ph.D. Student
Natural Computing Group (GCN)
Universidad Politécnica de Madrid
Spain

ATINER CONFERENCE PAPER SERIES No: COM2013-0595

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209
Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

Printed in Athens, Greece by the Athens Institute for Education and Research.

ISSN **2241-2891**
1/10/2013

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Guillén, P.G. (2013) **"Topological Proof of the Computability of the Algorithm based on the Morphosyntactic Distance"** Athens: ATINER'S Conference Paper Series, No: COM2013-0595.

Topological Proof of the Computability of the Algorithm based on the Morphosyntactic Distance

Pedro G. Guillén
Ph.D. Student
Natural Computing Group (GCN)
Universidad Politécnica de Madrid
Spain

Abstract

Following the previous works of E. Villa, A. De Santos and P. G. Guillén, considering a natural language it is possible to built the lexical associated space as a free semigroup, with the grammatic rules as its restrictions. Through several quotients and manipulations the univocal language is built, solving the problems of polisemy and synonymy that comes with the given natural language. Since here, the Morphosyntactic Distance can be defined over the elements of this group regardless of its algebraic properties, from a linguistic criterion. Therefore, it induces a topological space, which is called Morphosyntactic Space. Based on these hypothesis, some properties of this space are studied in this paper from a topological point of view, as compactness, total disconnection and separation. Later, the space is related through homeomorphisms, continuous functions and injective and surjective functions with some mathematical known spaces. After, a proof of the computability of the associated algorithm is given from these properties. Beyond the concrete problem which is solved in the paper with a topological argument, is shown that the method used in the proof could be generalized for an entire class of problems related with linear programming.

Keywords:

Corresponding Author:

Let be L a natural language. As in [1], we built the lexical space D, that haves a semigroup structure, and could be treated as a set, regardless of its algebraic properties. By [2], and using the fact that the meaning function is inyective, we are able to define in D as a set the Morphosyntactic Distance d, giving it a metric space structure. Thus, a topological space has been defined with the topology induced by d, that we wil call Morphosyntactic Topology, and the space (D,T) will be called Morphosyntactic Space. Trough this paper these hypothesis will be asumed as starting point. Analogue constructions of the same idea can be found in [7] and [8], since where we can assure that the chosen structure is consistent with a computational implementation.

Note that, because the topology is induced by a distance, the morphosyntactic space is metrizable. Therefore, it must apply the Metrization Theorem of Urysohnn to conclude that is a second-countable space and normal. To analize the connection properties, we introduce the following proposition.

Proposition 1. The Morphosyntactic Space is totally disconnected.

Proof. As can be seen in [2], the Morphosyntactic Space is countable, and considering x in D, there can not be any element y in D that verifies d(x,y)=0. Also by [2], we can asume that the set of values that d(x,y) can reach is bounded below for a constant k . Therefore, is clear that $B(x, k/2) \cap (D,T) = \{x\}$, from which it follows the result.

Proposition 2. The Morphosyntactic Space is compact.

Proof. As seen above, the Morphosyntactic Space is countable, so it must be trivially a Lindelöf space. Moreover, being a metrizable space is paracompact. Therefore, as Lindelöf and paracompact, the morphosyntactic space is compact.

Corollary 1. Morphosyntactic Space has a countable dense subset.

Proof. Being a compact and metrizable space, the result is immediate. A formal developement can be seen in [3].

From the reasoning carried out so far we have some basic properties of the Morphosyntactic Space. Our next step is define a similarity between this and a known structure that could constitute a reference in subsequent disquisitions. Related structures Hereinafter, we will make a comparison between different structures and the morphosyntactic space, in order to achieve a more intuitive approach to a future computational implementation of the model.

Theorem 1. The Morphosyntactic Space can be embedded in a Hilbert Cube.

Proof. We must prove that there is a continuous and injective function that $h:(D,T)\rightarrow[0,1]x[0,1]...$ By Corollary 1, a countable dense set $X = \{ a_1, a_2, \ldots\}$ can be taken in (D,T) to define the function $h(p) = (d(p,a_1), d(p,a_2), \ldots)$ Firstly, we will show the continuity of the function h. According with [3], we need to verify that for all natural number n, $\pi_n(h(x))$ is a continuous mapping. Let be n a natural number. Calculating $\pi_n(h(x)) = \pi_n(d(p,a_1),d(p,a_2),\ldots) = d(p,a_n)$. We must prove that the distance to a fixed element is a continous function. For this, let be the function defined as $H:(D,T) \rightarrow[0,1]$ that $H(x) = d(p,x)$ as $d(p,x) \leq d(x,y) + d(y,p)$ then $d(x,p) - d(y,p) \leq d(x,y)$. Similarly, $d(y,p) - d(x,p) \leq d(y,x)$. Therefore, we can conclude that $|H(x)-H(y)| \leq d(x,y)$ from it follows the

continuity. Since here, h is continous. We will show that h is injective. Proceeding by reduction ad absurdum, let suppose that there exists elements p,q in (D,T) that $p \neq q$ and h(p) = h(q). Since $p \neq q$, the number d(p,q) = c is positive. Since X is dense in (D,T), $B(p,c/2\}) \cap X \neq \{\emptyset\}$. So there exist a natural number r that a_r is in B(p,c/2). As we are supposing that h(p) = h(q), we have that $(d(p,a_1),d(p,a_2),..) = (d(q,a_1),d(q,a_2),\ldots)$ Then $d(p,a_n) = d(q,a_n)$ for all natural n. In particular, $d(p,a_r)=d(q,a_r)$, from we have $c = d(p,q) \leq d(p,a_r) + d(q,a_r) = 2d(p,a_r) < 2c/2 = c$ and c < c. ¡! Therefore, h is injective. And finally, analyzing the direct image of X, must be $h(X) \subseteq [0,1]x[0,1]x[0,1]x\ldots$ Moreover, h(X) is a compact (is the continuous image of a compact). We know that the compact subspaces of a metric space are closed then h(X) is closed.

Lemma 1. Let be C the Cantor set, and $A \subseteq C$ a nonempty closed set. Then, there exists a continous function k: $C \rightarrow A$ k(a) = a for all a in A.

Proof. As we know, $C \approx D$, where $D = \{\sum_{n=1}^{\infty} \frac{a_n}{4^n}$, $a_n \in \{0,3\}$ for all n$\}$. Let be k: $D \rightarrow A$ defined by $k(x) = a_x$, where $|x - a_x| = \min\{|x - a| : a \in A\}$. *Id est*, k(x) is the nearest point to x. Thinking about the distance as a function, we have seen above that this function is continous. Since A is a closed in a compact, then this function reaches minimum in an A. This tells us that there exists such an a_x. We must now prove that a_x is unique. Lets suppose that there are then a_x and b_x in A such that $|x - a_x| = |x - b_x| = \min\{|x - a|:a \text{ is in } A\}$. We may suppose that $a_x < b_x$. If $x \leq a_x < b_x$, then $|x - a_x| < |x - b_x|$ what is absurd. Similarly, $b_x \leq x$ can not ocurr. Then $a_x < x < b_x$ and $|x - a_x| = |x - b_x|$, what is also absurd. With this contradiction we prove that a_x is unique and therefore k is well defined. If x is in A, then x is the nearest point to x. Therefore, k(x) = x. We will show that k is a continuous function. Let be x in D and $\varepsilon > 0$. We will discuss fisrt the case where $x \notin A$. Lets suppose, without lose of generality, that $x < a_x$. Let be $\rho = a_x - x$, we will note $z = x - \rho$. Then $z < x < a_x$ and $|a_x - x| = |x - z|$. Clearly, is not possible that z, x, a_x in D. As x, a_x in D, we deduce that $z \notin D$. Being D a closed set, it must exist $r > 0$ that $(z - r, z + r) \cap D = \{\emptyset\}$ and $r < \rho$. We are going to prove now that $k(y) = a_x$ for all y in $(x - r/2, x + r/2) \cap D$. To do this, we have only to see that $|y - a_x| \leq |y - a|$ for all a in A. Let be then a in A. By definition of ρ, a_x we have that $\rho = |x - a_x| \leq |x - a|$. Thus, $(a \leq x - \rho) \vee (x + \rho \leq a)$ *id est*, $(a \leq z) \vee (a_x \leq a)$. Case 1: $a \leq z$. As $(z - r, z + r) \cap D = \{\emptyset\}$ and a in D, then $a \leq z - r$ or $z + r \leq a$. The second inequality is clearly false, because $a \leq z$. Therefore, $a \leq z - r$, and $a \leq z - r < z < x - r/2 < y$. So we have that $|y - a| = y - a > x - r/2 - (z - r) = (x - z) + r/2 = \rho + r/2 = a_x - x + r/2 = a_x - (x - r/2) > a_x - y = |a_x - y|$. Therefore, $|y - a_x| \leq |y - a|$. Case 2: $a_x \leq a$. In this case $|y - a_x| = a_x - y \leq a - y = |y - a|$. Consequently, $k(y) = a_x$ for all y in $(x - r/2, x + r/2) \cap D$. Taking $\delta = r/2$, we have that, for all y in $(x - r/2, x + r/2) \cap D$, $|k(y) - k(x)| = |a_x - a_x| = 0 < \varepsilon$. Thus, k is continuous in D-A. Lets discuss now the case where x in A. Then we take $\delta = \varepsilon/2$. If y is in $B(x,\delta) \cap D$, then by definition and as x is in A, we have that $|y - k(y)| = |y - a_y| \leq |x - y| + |y - k(y)| < \varepsilon$. Therefore, k is continuous in A.

Theorem 2. Let C be the Cantor set. Then there exists a continuous and surjective function $\phi:C\rightarrow(D,T)$

Proof. By Theorem 1, there exists a continous injective function $h:(D,T) \rightarrow[0,1]x[0,1]x\ldots$ Moreover, we know by [3] that there exists a continous and surjective function $f:C\rightarrow[0,1]x[0,1]x\ldots$ Let be $B = f^{-1}(h(D,T))$. B is clearly a nonempty closed set of C. By Lemma 1, there exists a continuous function $k:C\rightarrow B$ such that $k(b) = b$ for all b in B. Note that h is an homeomorphism over its image. We define now the function $\phi = h^{-1}(f(k)):C\rightarrow(D,T)$. ϕ is a continuous function, as being a composition of continuous functions. Given p in C, k(p) is in $B = f^{-1}(h(X))$. So f(k(p)) is in h(X). Since here, it has sense apply $h^{-1}(f(k(p)))$. This tells us that ϕ is well defined. To show that ϕ is suprajective, lets take an arbitrary element x in X. Let be $q = h(x)$ in h(X). As f is suprajective, there exists c in C such that $f(c) = q = h(x)$. Then c is in $f^{-1}(h(X)) = B$, and consequently $k(c) = c$. Therefore, $\phi(c) = h^{-1}(f(k(c))) = h^{-1}(f(c)) = h^{-1}(q) = x$. Thus, ϕ is suprajective.

Theorem 3. There exists an homeomorphism $f:(D,T)\rightarrow(G,T)$, where (G,T) is a graph with the discrete topology.

Proof. By Proposition 1, we can define rightly the function f identifing the elements x,y of (D,T) with the nodes of G and the number d(x,y) as the weight of the edge that connects them. Clearly, f is an homeomorphism.

Under the latter result, we can assume a reasonable time to implement search, filtering, and other type of algorithms into the Morphosyntactic Space. We can also simplify the development of computational tools associated with semantic spaces, following the example of [4] and [5].

This construction allows us to deduce the computability of the algorithm associated to find the distance between two given words in Morphosyntactic Space, because the algorithm is represented as a linear transformation on a finite dimensional space in which the elements have a locally unique binary representation. Therefore, the linear application itself can be represented in a finite binary sequence, being this trivially computable.

Moreover, the method described above allows us to generalize the demonstration of computability to all data sets that can be expressed as a continuous image of the Cantor set surjective, because both the array elements as described processes can be identified including applications linear, and these with finite dimension binary sequences.

References

Guillén, P. G., Villa, E., De Santos, A, Serradilla, F. (2012). 'Semantic Construction of an Univocal Language.' *Information Theories and applications*, 19(3): 211-215.

Guillén, P. G., Villa, E., De Santos, A., López Tolic, O. (2012). 'Construction of Morphosyntactic Distance on Semantic Structures.' *Information Theories and applications*, 19(4): 330-336.

Willard, S. (1970). *General Topology*. Reading, Massachusetts: Addison-Wesley.

Melton, A. (1989). 'Topological spaces for CPOS.' *Lecture Notes in Computer Science: Categorical Methods in Computer Science: with Aspects From Topology* 393: 302-314.

Weihrauch, K., Schreiber, U. (1981). 'Embedding metric spaces into CPO's.' *Theoretical Computer Science*, 16(1): 5-24.

Nicolov, N., Bontcheva, K., Angelova, G. and Mitkov, R. (2003). 'Recent advances in natural language processing, III.' Paper presented at the International Conference on Recent Advances in Natural Language Processing (RANLP 2003), September 10–12, in Samokov, Bulgaria.

Heckmann, R., Huth, M. (1998). 'Quantitative Semantics, Topology and Posibility Measures.' *Topology and Its Applications* 89: 151-178.

Blanck, J. (2000). 'Domain Representations of Topological Spaces.' *Theoretical Computer Science* 247(1-2): 229-255.