CONSTRUCTION OF MORPHOSYNTACTIC DISTANCE ON SEMANTIC STRUCTURES

Eduardo Villa, Alejandro De Santos, Pedro G. Guillén, Octavio López Tolic

Abstract: *When a natural language is introduced in a computer we have several problems to automate it, this is because it is difficult to make a computer relate concepts in a rich and always changing vocabulary.*
The two biggest problems come when using synonymous and polysemous words, as it is done in any natural language, this is the problem approached in this paper with an efficient algorithm.
The association of a natural language and a space based metric on semantic is probably the best way to manage any language with a machine, so building Δ as a freesemigroup with its grammatical rules as its natural restrictions, lets us define a distance that helps to provide a metric to the Morphosyntactic space so it is possible to organize and, therefore, study, in an automatic way, a natural language.
Keywords: *Natural Language, Semantic, Morphology, Distance, Syntactic, Context, Metric Space.*
ACM Classification Keywords: *D.2.11 [Software Engineering] Languages; E.1 [Data Structures] Graphs and networks; H.5.2 [User Interfaces] Natural language; I.7.0 [Document And Text Processing] General.*

Introduction

Let $\mathcal{L}$ be a natural language. Building the lexical space Δ, with the structure of a free semigroup, and which will be treated as a set, regardless of its algebraic properties relating its elements.
Let a word be treated as a pair (Θ, Ω) , data, context respectively, (see [Ito, 1977]) where a data is an array of letters with one or more meanings and a context as a set of words.
Starting with the hypothesis that a data has an only one meaning in a concrete context and,inspired to the idea of treating a word as a pair, the problem that polysemy represents is easily approached.
The synonymy problem is treated relating a word with its meaning so words with same meaning are the in the same group and, from now on, will be treated as a unique word.
Also the set of contexts will be treated as a finite, computable, arbitrary graph with contexts as nodes and natural relations between contexts as edges.
By the assumption of these hypothesis, and with the use of the fact that the meaning function described above, that associates each word with its meaning, is injective, it is possible to define in Δ as a set, the Morphosyntactic Distance $\boldsymbol{d}$ by comparing pairs (as in [Ito, 1981]), and to give it a metric space structure.
To start dealing with this problem some concepts may be defined.

Analysis of the Words

$\boldsymbol{Definition\ 1}$. Lexeme space $\mathcal{L}_e$.
The lexeme space,$\mathcal{L}_e$, is definedas the set formed by all the lexemes taken of each word of our language.
$\boldsymbol{Definition\ 2}$. Morpheme space $\boldsymbol{M}_e$.
The morpheme space, $\boldsymbol{M}_e$, is defined as the set formed by all the morphemes taken of each word of our language.
$\boldsymbol{Definition\ 3}$. Main space$\mathcal{M}$.
The main space, $\mathcal{M}$, is defined as

$$\mathcal{M} = \mathcal{L}_e \bigcup M_e$$

$\boldsymbol{Definition\ 4}$. Decomposition of a word.

The decomposition of a word $\boldsymbol{P}$ is defined asits only decomposition such that its morphemes and lexemes are separated for its right classification (see [Jacquemin, 2005] and [Koskenniemi, 1984]), taking similar principles as in [Karttunen, 1992](see below):

$$p_0 = (L^1, \dots, L^m, M^1, \dots, M^n)$$

with $\{L^1, \dots, L^m\} \in \boldsymbol{\mathcal{L}_e}$ and $\{M^1, \dots, M^n\} \in M_e$ so $p_0 \in < M >$.

Taking as principles:

$(\boldsymbol{i})$ Different forms of similar lexemes are mapped to the same canonical form.

$(\boldsymbol{ii})$ Morphological categories, such as plural, comparative or first person, are classified and compared between them.

E.g. $\boldsymbol{nonsmokers}$ is decomposed as $(\boldsymbol{non, smok, er, s})$.

$\boldsymbol{Observation\ 1}$. The main space is a countable space.

$\boldsymbol{Proof}$. The product of two countable sets is countable.

$\boldsymbol{Definition\ 5}$. Main lexeme of a word L^0.

The main lexeme of a word, in the case of compound words, is its most relevant lexeme in a determined context, being different in different cases, in the case of simple words it is its only lexeme.

The main lexeme of a word can be chosen as in [Jacquemin, 2001].

$\boldsymbol{Definition\ 6}$. Main decomposition of a word.

The main decomposition of a word $\boldsymbol{P}$ is a realignment of its main decomposition with its main lexeme as its first element:

$$p = (L^0, L^1, \dots, L^{m-1}, M^1, \dots, M^n)$$

E.g. $\boldsymbol{bullfrogs}$ is decomposed as $(\boldsymbol{bull, frog, s})$ and has as main decomposition $(\boldsymbol{frog, bull, s})$ using $\boldsymbol{frog}$ as its main lexeme.

In the next section an easy computable distance is defined so its elements can be treated as points in a metric space.

Morphosyntactic Distance

Now some useful operations between two elements generated by $\boldsymbol{\mathcal{M}}$denoted by $\Delta = < \boldsymbol{L'} > \subseteq L$, with its image in$\mathbb{R}$are defined. Let $\boldsymbol{P_1, P_2} \in (\Theta, \Omega)$ be, and considering its respective main decompositions:

$$p_1 = (L_1^0, L_1^1, \dots, L_1^{m-1}, M_1^1, \dots, M_1^n)$$
$$p_2 = (L_2^0, L_2^1, \dots, L_2^{m-1}, M_2^1, \dots, M_2^n)$$

Without loss of generality it can be supposed that $\boldsymbol{m + n \leq r + s}$.

Let $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon_0}, \boldsymbol{\varepsilon_1}, \dots, \boldsymbol{\varepsilon_{m+n}}\}$ be with $\boldsymbol{\varepsilon_i} \in \mathbb{Q} \cap (\boldsymbol{0}, \boldsymbol{1/2}) \ \forall \boldsymbol{i} \in \{\boldsymbol{0}, \dots, \boldsymbol{m + n}\}$ and $\boldsymbol{\varepsilon_0} > \prod_{\boldsymbol{i=1}}^{\boldsymbol{m+n}} \boldsymbol{\varepsilon_i}$.

Let $\boldsymbol{L_P} = (\boldsymbol{L^0, L^1, \dots, L^a})$ be.

Let $\boldsymbol{M_P} = (\boldsymbol{M^0, M^1, \dots, M^b})$ be.

Let the next algorithm be defined:

$$\delta_0 \colon (\Theta, \Omega) \mathrm{x} (\Theta, \Omega) \longrightarrow [0,1]$$

$\delta_0(P_1, P_2)$
if $p_1 = p_2$
    $\delta_0(P_1, P_2) = 0$
end
else
    $i = 0$
    $\delta_0(P_1, P_2) = 1$
    if $L_1^0 = L_2^0$
        $\delta_0(P_1, P_2) = {}^1/_2 + \varepsilon_0$

```
        end
        for L_1  in L_P1 + M_P1
              for  L_2  in L_P2 + M_P2
                    if L_1 = L_2
                          i = i + 1
                    end
              end
        end
        for j  in range (1, i)
              δ_0(P_1, P_2) = (δ_0(P_1, P_2) − 1/2)ε_j + 1/2
        end
end
```

$\boldsymbol{Observation\ 2}$. Let $\boldsymbol{p_1}, \boldsymbol{p_2}$ be the main decompositions of $P_1, P_2 \in \Theta$ such that they share s morphemes and non-main lexemes:

$$\delta_0(P_1, P_2) = \begin{cases} \varepsilon_0\varepsilon_1 \dots \varepsilon_s + 1/2 \ if\ L_1^0 = L_2^0 \\ \varepsilon_1 \dots \varepsilon_s + 1/2 \ L_1^0 \neq L_2^0 \end{cases}$$

$\boldsymbol{Example}$.

$\delta_0(P_1, P_2)$ with $P_1 = (bullfrogs, animal),\ P_2 = (frogmans, sea) \in (\Theta, \Omega)$

$$p_1 = (frog, bull, s), L_{P1} = (\text{frog, bull}), M_{P1} = (s)$$

$$p_2 = (man, frog, s), L_{P2} = (\text{man, frog}), M_{P2} = (s)$$

$$\delta_0((\text{bullfrogs, animal}), (\text{frogmans, sea}))$$

$frog \neq man \rightarrow \delta_0(P_1, P_2) = 1$

$$L_1^0 = L_2^1 \rightarrow \delta_0(P_1, P_2) = \varepsilon_1 + 1/2$$

$M_1^1 = M_2^1 \rightarrow \delta_0(P_1, P_2) = \varepsilon_1\varepsilon_2 + 1/2$

Function $\boldsymbol{\delta_1}$: (Θ, Ω)x$(\Theta, \Omega) \rightarrow [\mathbf{0}, \mathbf{1}]$ is defined as $\boldsymbol{\delta_1(P_1, P_2) = \min\{\delta_0([P_1], [P_2])}$ with $[\boldsymbol{P_1}], [\boldsymbol{P_2}]$ equivalence classes of $[\boldsymbol{P_1}]$ and $[\boldsymbol{P_2}]$respectively.

$\boldsymbol{Proposition\ 1}$. Function $\boldsymbol{\delta_1}$ keeps triangle inequality.

$\boldsymbol{Proof}$.

$$\boldsymbol{\delta_1(x, y) \leq \delta_1(x, z) + \delta_1(z, y)}$$

Several cases can be distinguished:

1.If $\boldsymbol{x = y}$ then $\boldsymbol{\delta_1(x, y) = 0}$ and $\boldsymbol{\delta_1(x, z) + \delta_1(z, y) \geq 0}$.

2.If $\boldsymbol{x \neq y}$ and $\boldsymbol{z = x}$ then $\boldsymbol{\delta_1(x, y) = \delta_1(x, z) + \delta_1(z, y)}$ because $\boldsymbol{\delta_1(x, z) = 0}$ and $\boldsymbol{\delta_1(x, y) = \delta_1(z, y)}$.

3. If $\boldsymbol{x \neq y}$ and $\boldsymbol{z = y}$then it can beproved analogously.

4.If $\boldsymbol{x \neq y,\ z \neq x}$ and $\boldsymbol{z \neq y}$ then $\boldsymbol{\delta_1(x, y) \leq 1}$ and $\boldsymbol{\delta_1(x, z) + \delta_1(z, y) \geq 1/2 + 1/2 = 1}$, so the inequality is verified.

We define function $\boldsymbol{\delta_2}$: Ωx$\Omega \rightarrow [\mathbf{0}, \mathbf{1}]$ as:

$$\delta_2(C_1, C_2) \begin{cases} 1 - 1/n + 1 \ if\ C_1 \neq C_2 \\ 0 \qquad\qquad if\ C_1 = C_2 \end{cases}$$

being $\boldsymbol{n}$ the natural distance between $C_1$and C_2 on a graph with all its links lengths $\mathbf{1}$.

$\boldsymbol{Proposition\ 2}$. Function δ_2 is a distance.

$\boldsymbol{Proof}$.

$$(i)\ \delta_2(C_1, C_2) \geq 0$$

$\boldsymbol{\delta_2(C_1, C_2) = 1 - {1}/{n+1}}$ with$\geq \mathbf{0}$, from where the inequality is proved trivially.

$$(ii)\delta_2(C_1, C_2) = 0 \leftrightarrow C_1 = C_2$$

$\boldsymbol{\delta_2(C_1, C_2) = 0 = 1 - {1}/{n+1} \leftrightarrow 1 = {1}/{n+1} \leftrightarrow n+1 = 1 \leftrightarrow n = 0 \leftrightarrow C_1 = C_2}$

$$(iii)\delta_2(C_1, C_2) = \delta_2(C_2, C_1)$$

As n is a distance, we can assert $\boldsymbol{\delta_2(C_1, C_2) = 1 - {1}/{n+1} = \delta_2(C_2, C_1)}$

$$(iv)\delta_2(C_1, C_2) \leq \delta_2(C_1, C_3) + \delta_2(C_3, C_2)$$

Several cases can be distinguished:

1. If $\boldsymbol{C_1 = C_2}$ then $\boldsymbol{\delta_2(C_1, C_2) = 0}$ and $\boldsymbol{\delta_2(C_1, C_3) + \delta_2(C_3, C_2) \geq 0}$.
2. If $\boldsymbol{C_1 \neq C_2}$and $\boldsymbol{C_3 = C_1}$ then $\boldsymbol{\delta_2(C_1, C_2) = \delta_2(C_1, C_3) + \delta_2(C_3, C_2)}$ because $\boldsymbol{\delta_2(C_1, C_3) = 0}$ and $\boldsymbol{\delta_2(C_1, C_2) = \delta_2(C_3, C_2)}$.
3. If $\boldsymbol{C_1 \neq C_2}$ and $\boldsymbol{C_3 = C_2}$ then it can be proved analogously.
4. If $\boldsymbol{C_1 \neq C_2}$, $\boldsymbol{C_3 \neq C_2}$ and $\boldsymbol{C_3 \neq C_2}$ then $\boldsymbol{\delta_2(C_1, C_2) = 1 - 1 / (n_1 + 1) \leq 1}$ and $\boldsymbol{\delta_2(C_1, C_3) + \delta_2(C_3, C_2) = 1 - {1}/{n_2+1} + 1 - 1 / (n_3 + 1) \geq 1}$, so the inequality is verified.

Function $\boldsymbol{d}$: ΔxΔ→ $\mathbb{R}$ is defined as $\boldsymbol{d(\Delta_1, \Delta_2) = \delta_1(P_1, P_2) + \delta_2(C_1, C_2)}$.

$\boldsymbol{Theorem\ 1}$. Function $\boldsymbol{d}$ is a distance on knowledgespaceΔ, this distance is called Morphosyntactic distance.

$\boldsymbol{Proof}$.

$$(i)d(\Delta_1, \Delta_2) \geq 0$$

$\boldsymbol{d(\Delta_1, \Delta_2) = \delta_1(P_1, P_2) + \delta_2(C_1, C_2)}$ with $\boldsymbol{\delta_1(P_1, P_2) \geq 0}$ and $\boldsymbol{\delta_2(C_1, C_2) \geq 0}$, from where the inequality is obtained trivially.

$$(ii)d(\Delta_1, \Delta_2) = 0 \leftrightarrow \Delta_1 = \Delta_2$$

$\Longrightarrow$

$\boldsymbol{d(\Delta_1, \Delta_2) = \delta_1(P_1, P_2) + \delta_2(C_1, C_2)}$, it is known that $\boldsymbol{\delta_1(P_1, P_2) \geq 0}$ and $\boldsymbol{\delta_2(C_1, C_2) \geq 0}$, so it can be deduce that $\boldsymbol{\delta_1(P_1, P_2) = 0}$ and $\boldsymbol{\delta_2(C_1, C_2) = 0}$ what implies, because of the construction of $\boldsymbol{\delta_1}$ and$\boldsymbol{\delta_2}$, that $\boldsymbol{P_1 = P_2}$ and$\boldsymbol{C_1 = C_2}$, thus it is concluded that $\Delta_1 = (P_1, C_1) = (P_2, C_2) = \Delta_2$.

The other implication can be equally proved.

$$(iii)\boldsymbol{d(\Delta_1, \Delta_2) = d(\Delta_2, \Delta_1)}$$

As function $\boldsymbol{\delta_1}$is a combination of boolean comparisons to which apply an algorithm is applied and these comparisons do not depend on the order of its elements, it can be inferredtrivially that $\boldsymbol{\delta_1(P_1, P_2) = \delta_1(P_2, P_1)}$ and, as $\boldsymbol{\delta_2}$ is a distance, $\boldsymbol{\delta_2(C_1, C_2) = \delta_2(C_2, C_1)}$, because all these facts, it can beassumed that

$$\boldsymbol{d(\Delta_1, \Delta_2) = \delta_1(P_1, P_2) + \delta_2(C_1, C_2) = \delta_1(P_2, P_1) + \delta_2(C_2, C_1) = d(\Delta_2, \Delta_1)}$$

$$(iv)\boldsymbol{d(\Delta_1, \Delta_2) \leq d(\Delta_1, \Delta_3) + d(\Delta_3, \Delta_2)}$$

As $\boldsymbol{\delta_1(P_1, P_2) \leq \delta_1(P_1, P_3) + \delta_1(P_3, P_2)}$and $\boldsymbol{\delta_2(C_1, C_2) \leq \delta_2(C_1, C_3) + \delta_2(C_3, C_2)}$ it can be concluded that

$\boldsymbol{d(\Delta_1, \Delta_2) = \delta_1(P_1, P_2) + \delta_2(C_1, C_2) \leq \delta_1(P_1, P_3) + \delta_1(P_3, P_2) + \delta_2(C_1, C_3) + \delta_2(C_3, C_2)}$.

$\boldsymbol{Proposition\ 3}$. Distance $\boldsymbol{d}$ between two different elements, $\Delta_1, \Delta_2 \in \Delta$ is bounded both, from above and below by ${1}/{2}$ and **2** respectively, $\boldsymbol{{1}/{2} \leq d(\Delta_1, \Delta_2) \leq 2}$.

$\boldsymbol{Proof}$.

It can besupposed, without loss of generality, that they share $\boldsymbol{s}$ morphemes and non-main lexemes and its contexts natural distance is $n \in \mathbb{N}$.

$$\boldsymbol{{1}/{2} = {1}/{2} + 1 * 0 + 1 - \frac{1}{0+1} \leq {1}/{2} + 1 * \varepsilon_0\varepsilon_1 \ldots \varepsilon_s + 1 - \frac{1}{n+1} = \delta_1(P_1, P_2) + \delta_2(C_1, C_2)}$$

$$= \boldsymbol{d}(\Delta_1, \Delta_2) = \boldsymbol{\delta_1}(\boldsymbol{P_1}, \boldsymbol{P_2}) + \boldsymbol{\delta_2}(\boldsymbol{C_1}, \boldsymbol{C_2}) = {}^{\mathbf{1}}\!/_{\mathbf{2}} + \mathbf{1} * \boldsymbol{\varepsilon_0}\boldsymbol{\varepsilon_1} \dots \boldsymbol{\varepsilon_s} + \mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}+\mathbf{1}} \leq \mathbf{1} + \mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}+\mathbf{1}} \leq \mathbf{2}.$$

Conclusion

With all these bases a topological space is built in (Θ, Ω) withthe topology induced by the Morphosyntactic distance.

Further study of it can give possibilities of working easily and faster with it, and bringing the opportunity to come to a more general algorithm to compare sentences and even full texts.

With this idea it is possible to relate different information and have a better approach to organize it for its quick accessibility and make easier its sharing.

Acknowledgements

This work is supported byAXON Ingeniería y Desarrollo de Software, S.L. (Spain); Buaala.TV Project,Avanza 2 TSI-090302-2011-19;Ministerio de Industria, Turismo y Comercio(Spain) and FEDER (European Union)

Bibliography

[Ito, 1977] Ito, Tetsuro; Kizawa, Makoto.Semantic structure of natural language.Systems-Computers-Controls 7 (1976), no. 2, 1–10 (1977).

[Ito, 1981] Ito, Tetsuro; Toyoda, Junichi; Kizawa, Makoto, Hierarchical data base organization for document information retrieval.Systems-Comput.-Controls 10 (1979), no. 2, 39–47 (1981).

[Jacquemin, 2001]Jacquemin, Christian .Spotting and Discovering Terms through Natural Language Processing. 2001. The MIT Press. Page 16.

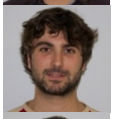
[Jacquemin, 2005]Jacquemin, Christian .Spotting and Discovering Terms through Natural Language Processing. 2001. The MIT Press. Page 16.Creutz, Mathias.Lagus, Krista. Inducing the Morphological Lexicon of a Natural Language from Unannotated. 2005. International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning. Espoo. Finland.

[Karttunen, 1992]Karttunen, Lauri; Kaplan, Ronald M.; Zaenen, Annie.Two-level morphology with composition. 1992. COLING '92 Proceedings of the 14th conference on Computational linguistics - Volume 1. Stroudsburg, PA, USA.

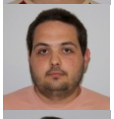
[Koskenniemi, 1984]Koskenniemi, Kimmo. A general computational model for word-form recognition and production. 1984. ACL '84 Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics. Stroudsburg, PA, USA.

Authors' Information

Eduardo Villa *– Natural Computing Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: e.villa.valdes@gmail.com*



Alejandro De Santos *– Natural Computing Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: matematicofeliz@gmail.com*



Pedro G. Guillén *– Natural Computing Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: pedrogguillen@gmail.com*

Octavio López Tolic *–Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Carretera de Valencia Km. 7, 28031 Madrid; e-mail: oa@eui.upm.es*