SEMANTIC CONSTRUCTION OF UNIVOCAL LANGUAGE

Alejandro De Santos, Pedro G. Guillén, Eduardo Villa, Francisco Serradilla.

***Abstract**: In this paper a solution is propose to organize the space of words that exist in a specific language in their different semantic categories. By taking a natural language, we're going to define a unique meaning for each word, as a construction made (d,C) of pairs of words and contexts. On the other hand, let us consider the space of meanings. All the words that share meaning (synonymous words) can be associated with one meaning. This permitus to make a partition of the space of words in groups of synonyms. Finally, a classification of the space of words will be obtained in the different groups of words that share meaning. This allow to choose the useful word that represent a meaning, and reduce the number of words selecting one representative from each group of synonyms. It will be very useful for calculating distances between words.*

***Keywords**: Natural Language, Semantic, Context.*

***ACM Classification Keywords:** C.2.1 Network Architecture and Design; D.2.1 Requirements Specifications (D.3.1) Languages; D.3.2 Language Classifications Applicative (functional) languages; E.1 [Data Structures] Graphs and networks; H.5.2 [User Interfaces] Natural language.*

Introduction

Let A' be a natural language. We built $D = (A',*)$, with a free semigroup structure, treated as a set, regardless of its algebraic properties [Dieter, 2004]. Can be built the pair (d, C) made of one word and one context that select one of the possible meanings of the word. So, a solution to the problem of poly semy is proposed. Now, is necessary to consider the space Δ of all the possible meanings that built a language. Over the pair (d, C) the meaning mapping $\psi: (d, C) \longrightarrow \Delta$ is defined (we're going to include programming abstractions such as procedures, functions [Miguel, 2011]) ψ assigns to each pair (d, C) only one meaning. Finally, an equivalency relation over (d, C) is defined in order to build a quotient space inducing a classification in the different semantic classes [Ito, 1977], [Angelova, 1988], wherein each equivalence class is formed by all the words associated with a determinate meaning (synonymous words) [Ito, 1981].

Space of Words

$\boldsymbol{Definition}$ **1.** Let $\mathbf{A} = \{a, b, c \dots\}$ be an alphabet, such that $A \neq \emptyset$.

We can define a word as a finite succession of elements of $\boldsymbol{A}$, where repetitions are allowed.

For example **aaa**, **aabcc**, **b** are different words.

Let $\boldsymbol{A'}$ be the set composed by words.

$\boldsymbol{Definition}$ **2.** Let $\boldsymbol{D} = (\boldsymbol{A'},*)$ be a free semigroup with a associative law (*) juxtaposition, and an identity element $1 \in G$.

We can multiply two words **aaabbb*abccc=aaabbbabccc**, where the identity element '1' is defined as the empty succession of words.

The grammatical rules are defined as restrictions on this free semigroup.

$\boldsymbol{Definition\ 3}$. Let the inclusion mapping be $\boldsymbol{i_{|\Theta}}: \boldsymbol{D} \longrightarrow \boldsymbol{\Theta}$,Θ the space of concepts which belongs to a given lexicon $\mathcal{L}$. The inclusion mapping $\boldsymbol{i_{|\Theta}}$ saves the words that belong to $\boldsymbol{\mathcal{L}}$.

Let $d \in D$ then $\boldsymbol{i_{|\Theta}}(d) = \{d/d \in \Theta\}$ and $\boldsymbol{i_{|\Theta}}(d) = \{\emptyset/d \in D \backslash \Theta\}$.

For example $\boldsymbol{i_{|\Theta}}(\boldsymbol{aabbacc}) = \emptyset$ such that $\boldsymbol{aabbacc} \notin \boldsymbol{\mathcal{L}}$ and $\boldsymbol{i_{|\Theta}}(\boldsymbol{dog}) = \boldsymbol{dog}$ such that $\boldsymbol{dog} \in \boldsymbol{\mathcal{L}}.$

In the next point the space of context is building as a simple generation rule so as we will see how to calculate the shortest path between two contexts.

Space of Contexts

Let $\boldsymbol{C}$ be a context, a list of words that define an ambient where to locate a word belonging to $\boldsymbol{\mathcal{L}}$.

$\boldsymbol{Definition\ 4}$. Let $\boldsymbol{C}$ a graph made of $\{\boldsymbol{C_1}, \boldsymbol{C_2}, \dots \boldsymbol{C_n}\}$ a countable set of contexts, such that $\boldsymbol{C_i} \in \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ is the space of contexts inspired by [Dieter, 2004].

We're going to build C with a simple $\boldsymbol{formation\ rule}$. Let $\boldsymbol{C_i} \in \boldsymbol{\Omega}$ be a determined context composed by more specific disjoint contexts $\boldsymbol{C_{i+1}} \in \boldsymbol{\Omega}$.

$$C_i = \bigcup_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{C_{i+1}^{j}} \quad such\ as \cap_{j=1}^{\boldsymbol{n}} \boldsymbol{C_{i+1}^{j}} = \emptyset \equiv [4.1]$$

$\boldsymbol{Definition\ 5}$.Let the $\boldsymbol{generation\ rule}$ be the process of division of a context $\boldsymbol{C_i} \in \boldsymbol{\Omega}$ in their constituents subcontexts $\boldsymbol{C_{i+1}} \in \boldsymbol{\Omega}$ in $[4.1]$. The number of generation rules must be countable to avoid problems of computability.

We consider $\boldsymbol{C_i}$ as a vertex of the graph and all the contexts in which $\boldsymbol{C_i}$ is subdivided as a set $\{\boldsymbol{C_{i+1}^1}, \boldsymbol{C_{i+1}^2}, \dots \boldsymbol{C_{i+1}^n}\}$ of new vertex's connected with $\boldsymbol{C_i}$ by edges.

If we are applying the rule generation to $\boldsymbol{C_i}$ and one of the subcontext generated $\boldsymbol{C_{i+1}^3}$ is the same as another $\boldsymbol{C_{j+2}^5}$ belonging to another division $\boldsymbol{C_{j+2}^5} = \boldsymbol{C_{i+1}^3}$, then this two subcontexts are considered the same and we can see it in the graph using the edges from the previous settings $\boldsymbol{C_i}, \boldsymbol{C_{j+1}}$ to it.

This breaks the tree structure of the graph, creating closed paths and cycles.

Let $\|\boldsymbol{\alpha}(C_i, C_j)\|$ be the $\boldsymbol{lenght\ of\ a\ path}$ between two contexts [Dieter, 2004].

$\boldsymbol{Definition\ 6}$. The $\boldsymbol{shortest\ lenght\ of\ a\ path\ between\ two\ contexts}$ is:

$$\|\boldsymbol{\alpha_{min}}(C_i, C_j)\| = \{\|\boldsymbol{\alpha}(C_i, C_j)\| \ such\ that \forall\ \|\boldsymbol{\alpha}'(C_i, C_j)\|, \|\boldsymbol{\alpha}'(C_i, C_j)\| \geq \|\boldsymbol{\alpha}(C_i, C_j)\|\}$$

It will be useful to calculate in the future, the minimum distance between two contexts.

In the next point we're going to build the relationship between the space of words $\boldsymbol{\Theta}$ and space of contexts $\boldsymbol{\Omega}$ in the pair $(\boldsymbol{d}, \boldsymbol{C}) \in (\boldsymbol{\Theta}, \boldsymbol{\Omega})$.

Once this is achieved, we will assign one meaning to each pair through a well defined mapping.

The Meaning Mapping

$\boldsymbol{Lemma\ 1}$. Let $(\boldsymbol{d}, \boldsymbol{C}) \in (\boldsymbol{\Theta}, \boldsymbol{\Omega})$ be the pair made of one word and one context. $\boldsymbol{\Theta}$ is the space of concepts which belongs to a given lexicon, and $\boldsymbol{\Omega}$ is the space of contexts.

It is important to realize that most of the words of a given lexicon $\boldsymbol{\mathcal{L}}$ are polysemous.

Is selected one of the possible meanings of the word through the context where to locate the word.

Let $\boldsymbol{\Delta}$ be meanings space of a given lexicon $\boldsymbol{\mathcal{L}}$.

$\boldsymbol{Lemma\ 2.}$ Let $\boldsymbol{\psi}$: $(\boldsymbol{\Theta}, \boldsymbol{\Omega}) \longrightarrow \boldsymbol{\Delta}$ a mapping. $\boldsymbol{\Delta}$ is the meanings space, $\boldsymbol{\Theta}$ is the space of words that belong to $\boldsymbol{\mathcal{L}}$, and $\boldsymbol{\Omega}$ the space of contexts.

Once we have selected a word and a context by the pair, we assign one meaning to the pair through the mapping $\boldsymbol{\psi}$.

For example, the word “paint”, choose different meanings for different contexts through the mapping $\boldsymbol{\psi}$.

$\boldsymbol{\psi}(\boldsymbol{paint}, \boldsymbol{decoration})$ =a coloured substance which is spread over a surface and dries to leave a thin decorative or protective coating.

$\boldsymbol{\psi}(\boldsymbol{paint}, \boldsymbol{basketball})$ =a rectangular area marked near the basket at each end of a court.

$\boldsymbol{Proposition\ 1.}\ \boldsymbol{\psi}$ is a well-defined mapping.

$\boldsymbol{Proof.}$ By making $\boldsymbol{\psi}$ a mapping, we have found a way to avoid the polysemy problem and assign one only meaning to each pair (**word, context**) [Kazimierz, 2010].

Let $\boldsymbol{d} \in \boldsymbol{\Theta}$ be a word.

Let $\boldsymbol{C_1}, \boldsymbol{C_2}, \dots \boldsymbol{C_n}$ be a list of contexts associated to the word d.

$$\forall\ (\boldsymbol{d}, \boldsymbol{C_i})\ \exists!\ h\ \in \Delta /\ \boldsymbol{\psi}(\boldsymbol{d}, \boldsymbol{C_i}) = h \text{ with } i \in \{1,2, \dots n\}.$$

We can consider a pair $(\boldsymbol{d}, \boldsymbol{C}) \in (\boldsymbol{\Theta}, \boldsymbol{\Omega})$, the support of $\boldsymbol{\psi}$, that allow to define $\boldsymbol{\psi}$ as a mapping.

$\boldsymbol{\psi}$: $(\boldsymbol{\Theta}, \boldsymbol{\Omega}) \longrightarrow \boldsymbol{\Delta}$ defines only one meaning for each pair.

In the last point we will organize the pairs $(\boldsymbol{d}, \boldsymbol{C}) \in (\boldsymbol{\Theta}, \boldsymbol{\Omega})$ by groups that share meaning (synonymous words) through an equivalence relation $\sim$. This allow to choose the word that represents a meaning that suits us, and reduce the number of words by choosing one representative from each group of synonyms. It will also be very useful for calculating distances between words.

The space of semantic meanings

$\boldsymbol{Lemma\ 3.}$ Let $\sim$ be an equivalence relation. Two words in one of their concrete contexts are related: $(\boldsymbol{d}, \boldsymbol{C}) \sim (\boldsymbol{d}', \boldsymbol{C}')$ if and only if $\boldsymbol{\psi}(\boldsymbol{d}, \boldsymbol{C}) = \boldsymbol{\psi}(\boldsymbol{d}', \boldsymbol{C}')$.

For example:

$\boldsymbol{\psi}(\boldsymbol{late}, \boldsymbol{time})$ = after the expected or usual time.

$\boldsymbol{\psi}(\boldsymbol{delayed}, \boldsymbol{person})$ = retarded, incapacitated.

$\boldsymbol{\psi}(\boldsymbol{delayed}, \boldsymbol{traffic})$ = after the expected or usual time.

Where $\boldsymbol{\psi}(\boldsymbol{late}, \boldsymbol{time}) = \boldsymbol{\psi}(\boldsymbol{delayed}, \boldsymbol{traffic})$.

$\boldsymbol{Teorem\ 1.}$ $\sim$ is a **equivalence relation**.

$\boldsymbol{Proof.}$

$\sim$ is a **reflexive relation**: One word in one of its meanings is related with himself

$(\boldsymbol{d}, \boldsymbol{C}) \sim (\boldsymbol{d}', \boldsymbol{C}') \leftrightarrow \boldsymbol{\psi}(\boldsymbol{d}, \boldsymbol{C}) = \boldsymbol{\psi}(\boldsymbol{d}', \boldsymbol{C}')$

Obviously: $(\boldsymbol{late}, \boldsymbol{time}) \sim (\boldsymbol{late}, \boldsymbol{time}) \leftrightarrow \boldsymbol{\psi}(\boldsymbol{late}, \boldsymbol{time}) = \boldsymbol{\psi}(\boldsymbol{late}, \boldsymbol{time})$ =after the expected or usual time.

$\sim$ is a **symmetric relation**: If one word, in one of their meanings, is related with another one (in one of its meanings). This involves that the another word is related with the first word, each one, in one of its meanings.

$(\boldsymbol{d},\boldsymbol{C}) \sim (\boldsymbol{d}',\boldsymbol{C}') \leftrightarrow \boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C}) = \boldsymbol{\psi}(\boldsymbol{d}',\boldsymbol{C}') \equiv \boldsymbol{\psi}(\boldsymbol{d}',\boldsymbol{C}') = \boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C}) \leftrightarrow (\boldsymbol{d}',\boldsymbol{C}') \sim (\boldsymbol{d},\boldsymbol{C})$

Is not difficult to see:

$(\boldsymbol{late},\boldsymbol{time})\sim(d\boldsymbol{elayed},\boldsymbol{traffic}) \leftrightarrow \boldsymbol{\psi}(\boldsymbol{late},\boldsymbol{time})\sim\boldsymbol{\psi}(\boldsymbol{delayed},\boldsymbol{traffic}) =$ after the expected or usual time$=$ $\boldsymbol{\psi}(\boldsymbol{delayed},\boldsymbol{traffic}) = \boldsymbol{\psi}(l\boldsymbol{ate},\boldsymbol{time}) \leftrightarrow (\boldsymbol{delayed},\boldsymbol{traffic})\sim(\boldsymbol{late},\boldsymbol{time})$.

$\sim$ is a **transitive relation**: When one word(always in one of their meanings) is related with another and the third word is related with another one, this implies that the firs and the last word are related.

$$(\boldsymbol{d},\boldsymbol{C}) \sim (\boldsymbol{d}',\boldsymbol{C}') \text{ and } (\boldsymbol{d}',\boldsymbol{C}') \sim (\boldsymbol{d}'',\boldsymbol{C}'') \leftrightarrow \boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C}) = \boldsymbol{\psi}(\boldsymbol{d}',\boldsymbol{C}') = \boldsymbol{\psi}(\boldsymbol{d}'',\boldsymbol{C}'')$$

In this case:

$(\boldsymbol{late},\boldsymbol{time})\sim(\boldsymbol{delayed},\boldsymbol{traffic})$

and $(\boldsymbol{delayed},\boldsymbol{traffic})\sim(\boldsymbol{overdue},\boldsymbol{born}) \leftrightarrow \boldsymbol{\psi}(\boldsymbol{late},\boldsymbol{time}) = \boldsymbol{\psi}(\boldsymbol{delayed},\boldsymbol{traffic}) =$ $\psi(\boldsymbol{overdue},\boldsymbol{born})$that means as we know: after the expected or usual time.

This **equivalence relation** $\sim$ allow to organize the different pairs in their classes:

$[(\boldsymbol{d},\boldsymbol{C})] = \{(d',C') \in (\Theta,\Omega)/(\boldsymbol{d}',\boldsymbol{C}')\sim(\boldsymbol{d},\boldsymbol{C})\} \equiv \{(d',C') \in (\Theta\times\Omega)/\boldsymbol{\psi}(\boldsymbol{d}',\boldsymbol{C}') = \boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C})\}$

consisting of all possible pairs, formed by words and their contexts that share the same meaning.

This **equivalence relation** $\sim$ induces in $(\boldsymbol{\Theta}\times\boldsymbol{\Omega})$ the classification of their elements in the different semantic classes, building the quotient space $(\boldsymbol{\Theta},\boldsymbol{\Omega})/\sim\, = \{[(\boldsymbol{d},\boldsymbol{C})]/(\boldsymbol{d},\boldsymbol{C}) \in (\boldsymbol{\Theta}\times\boldsymbol{\Omega})\}$ wherein each element is one of the classes that share meaning.

For example:

$[(\boldsymbol{late},\boldsymbol{time})] = \{(\boldsymbol{delayed},\boldsymbol{traffic}) \in (\Theta,\Omega)/(\boldsymbol{late},\boldsymbol{time})\sim(\boldsymbol{delayed},\boldsymbol{traffic})\}$.

We can see that:

$$\cup_{\forall\, i,j}\left[\left(\boldsymbol{d}_i,\boldsymbol{C}_j\right)\right] = (\Theta\times\Omega)/\sim \text{ such that } \mathrm{j} \in \{1{,}2 \ldots \mathrm{n}\}, \mathrm{i} \in \{1{,}2 \ldots \mathrm{m}\}$$

$$\cap_{\forall\, i,j}\left[\left(\boldsymbol{d}_i,\boldsymbol{C}_j\right)\right] = \emptyset \text{ such that } \mathrm{j} \in \{1{,}2 \ldots \mathrm{n}\}, \mathrm{i} \in \{1{,}2 \ldots \mathrm{m}\}$$

Lemma4. Let $\boldsymbol{\psi}*$ be the meaning mapping defined on the quotient space$\boldsymbol{\psi}*\colon (\boldsymbol{\Theta},\boldsymbol{\Omega})/\sim\, \longrightarrow \boldsymbol{\Delta}$

$\boldsymbol{\psi}*$ is a mapping from the classes to the meanings

$$\boldsymbol{\psi}*[(\boldsymbol{late},\boldsymbol{time})] = \boldsymbol{\psi}(l\boldsymbol{ate},\boldsymbol{time}) = \boldsymbol{\psi}(\boldsymbol{delayed},\boldsymbol{traffic}) = \boldsymbol{\psi}(\boldsymbol{overdue},\boldsymbol{born})$$

that means as we know: after the expected or usual time.

Theorem2. The equivalent relation$\sim$makes be$\boldsymbol{\psi}*$a injective mapping.

Proof. $\boldsymbol{\psi}*[(\boldsymbol{d},\boldsymbol{C})] = \boldsymbol{\psi}*[(\boldsymbol{d}',\boldsymbol{C}')] \leftrightarrow [(\boldsymbol{d},\boldsymbol{C})] = [(\boldsymbol{d}',\boldsymbol{C}')] \equiv (\boldsymbol{d},\boldsymbol{C}) \sim (\boldsymbol{d}',\boldsymbol{C}')$

Two classes have the same meaning if they are the same class.

Finally the injective mapping $\boldsymbol{\psi}$ allows us to choose only one word that represents a meaning that suits us from each group of synonyms, simplifying the process.

We can see that:

$\boldsymbol{\psi}*[(\boldsymbol{late},\boldsymbol{time})] = \boldsymbol{\psi}*(\boldsymbol{delayed},\boldsymbol{traffic}) =$ after the expected or usual time.

Conclusion

We approach a solution to the problem of polysemy, building the different pairs $(\boldsymbol{d},\boldsymbol{C})$, choosing an unique meaning for each word. Acting on this pairs $\boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C})$, the meaning mapping $\boldsymbol{\psi}$, assigns one meaning to each pair. Finally, we have solved the synonymy problem, organizing all the words in their different semantic classes, through the quotient $(\boldsymbol{\Theta},\boldsymbol{\Omega})/\sim$**,** where two pairs $(\boldsymbol{d},\boldsymbol{C}) \sim (\boldsymbol{d}',\boldsymbol{C}')$ are related if and only if they share the meaning $\boldsymbol{\psi}(\boldsymbol{d},\boldsymbol{C}) = \boldsymbol{\psi}(\boldsymbol{d}',\boldsymbol{C}')$. With this action, we reduce the number of words by choosing one representative from each group of synonyms, allowing the selection of the word that represents the useful meaning.

Acknowledgements

This work is supported by AXON 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

[Angelova, 1988] Angelova, GalyaMladenova. Syntactic and semantic structures for relational query languages . (Russian) Tanulmányok-MTA Számítástech. Automat.Kutató Int. Budapest No. 209 (1988), 187 pp. 68P15 (68Q55)

[Dieter,2004] Dieter Jungnickel. Graphs, Networks and Algorithms. Volume5. Ed. Springer, 2004

[Ito, 1977] Ito, Tetsuro; Kizawa, Makoto. Semantic structure of naturallanguage. Systems-Computers-Controls 7 (1976), no. 2, 110 (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).

[Kazimierz,2010] Kazimierz Subieta. Syntax and Semantics of the Stack Based Query Language (SBQL)1, Polish-Japanese.Institute of Computer Science Polish Academy of Sciences Version 2, 18 June 2010. Page 41. Store mode Classes and inheritance.

[Miguel, 2011] Miguel Gonzalez. ESTRUCTURAS ALGEBRAICAS. 4de febrero de 2011 .Grupos libres. Generadores y relaciones. Página 93

Authors' Information

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*

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*

Francisco Serradilla *–MercatorGroup, Universidad Politécnica de Madrid Carretera de Valencia Km. 7, 28031 Madrid, Spain; e-mail: fserra@eui.es*