Modeling Graduality in Computerized Research on Synonyms

Farida REDJAI

LIP6 (ex. LAFORIA)-CNRS, Université Paris VI
Tour 46-00 2ème étage
4, Place Jussieu 75252 Paris Cedex 05, France
E.Mail: redjai@laforia.ibp.fr

Abstract

This paper deals with the problem of synonymy in linguistics and tries to bring a solution integrating the two notions of graduality and point of view in the automatic research on synonyms. The renunciation of the strict definition of synonymy which consists in interchanging two terms in all contexts allows us to sketch out a solution considering only the synonyms with respect to a given point of view.

The dictionary of synonyms, as we conceive it, is a domain-structured dictionary. In fact, it is obvious that a user who is looking for the synonyms of a word, already knows the domain, and wants to find the synonyms of this word precisely in this domain.

Our system constructs a "Domain-Fields-Taxemes" tree as a function of differential sema given by the expert. The leaves of this tree are taxemes and contain some words which, may be considered as equivalent with respect to one or several points of view.

In this way, we can avoid the often circular and infinite lists used in the current synonym dictionaries, and which may contain words which have nothing in common with the initial word. Furthermore, the use of many-valued logic has enabled us to present to the user an ordered list of synonyms, with the best synonym at the top of this list.

Introduction

In some contexts, it may not be possible to use indifferently a synonym instead of another. Apart from the common meaning that synonyms share, each synonym can have a specific connotation. Moreover, do the words "sword", "blade", "sabre" and "glaive" have exactly the same meaning? Even though they all relate to a common notion, they do carry some different connotations, they don't apply to identical objects, and they are not always used in an equivalent manner in the same context.

Differential semantics, with the help of semic analysis, allows us to decompose a word in a set of meanings called sema.

The synonym dictionary, as we conceive it, is structured into domains: it's obvious that when a user is looking for the synonyms of a word, he knows the *domain* he's working in, and wants to find synonyms precisely in that *domain.*

The system builds a "Domain-Fields-Taxemes" tree, as a function of *differential sema* that were given by the expert. A *field* is created for every differential *seme*. The leaves of this tree are *taxemes*, and contain words that can be considered as equivalent, with respect to one (or several) *point(s) of view* defined by the expert.

The first part of this paper is concerned with the *differential semantics*; the second part deals with *many-valued logics*, and the third one presents our solution for the problem of synonymy based on these two theories.

Differential semantics

The differential semantics which has been used in the framework of this study, offers a description of the word which does not correspond to a decomposition of the word into semantical primitives often conceived as semantic atoms with universal value. On the contrary, the differential features, or sema correspond to a synthetic description of the word and reject both the compositionality of the meaning and the existence of primitives independent of any context.

Semic analysis really took form in the early 60's (Pottier 1962), (Katz and Fodor 1963). It consists in enumerating the distinctive elements of *sememes* (specific and generic features).

A *sememe* is a set of pertinent features, also called sema. A *seme* is a minimal distinctive feature; it is a partial definition of a word.

Sema are based on opposition relations and equivalence relations between *sememes*. For instance, "bistoury" is opposed to "scalpel" by the seme /for alive creatures/.

However, two approaches are grouped under the term of semic analysis. The first is the referential approach (Katz 1972) that is based on the semasiological method, which is typical of lexicography. The sememes are analyzed according to their differences inside a class constituted by an expression criterion. For instance, if we regroup in a same class the sememes having for expression *"plane"*, we necessarily have to inter-define *"plane1"* (aeroplane) and *"plane2"* (carpenter tool), which have the same spelling. The second approach is the differential approach (Pottier 1962), (Coseriu 1976). It adopts an onomasiological method. The sememes are analyzed according to their differences inside a minimal semantic class (the *taxeme*), constucted from one strictly semantical criterion. For example, the class of *"furniture"* and the class of *"transportation means"*

For *differential semantics*, the number and the nature of the components of a sememe are directly related to the number and the nature of the components of the other sememes contained in its definition class.

On the contrary, for the referential semantics, there's no linguistic criterion which allows to choose the components, nor to limit their number. For instance, the *"chair"* is defined by 4 features in (Pottier 1974), and by 10 features in (Katz 1972).

Any semic analysis must of course solve the crucial problems of identification of features, and the limitation of their number.

Only the method of differential analysis can solve the problem of relevance of the components, since it operates on classes of words built according to linguistic criteria, but not on isolated words, in order to define them relatively to their referents.

The many-valued logics

The fundamental contribution of the *many-valued logics* and of the multisets theory lies in the introduction of truth-degrees to which correspond membership degrees (going from full-belonging to classical non-belonging) : a proposition may be neither true nor false, but true to a certain degree (Akdag 1992).

The truth-degree of such a many-valued proposition, or the degree of satisfaction of a property by an object, then coincides with the belonging degree of this objet to the corresponding multiset.

We present a model of knowledge representation inspired by the *many-valued logic* developed by De Glas (De Glas 1984).using a finite number of totally ordered truth-degrees which have a least element "not at all" denoted τ_1 and a greatest element "totally" denoted τ_M.

If we say :

"x is μ_α A"

that means "x (is τ_α) A"

that is: τ_α is a truth-degree, to which x is A.

In other words:

"x is μ_α A" is true

iff "x is A" is τ_α-true

Algebrical structure

The *many-valued logic* is a generalization of the Boolean logic. It uses a finite and totally ordered De Morgan lattice $L_M = \{\tau_1,\tau_2,\dots\dots\dots,\tau_M\}$ of truth-degrees ($\tau_i \leq \tau_j$ iff $i \leq j$) going from τ_1 ("not at all") to τ_M ("totally"), supplied with the $\vee$, $\wedge$, and ~ operators (Akdag 1992), (Akdag, De Glas and Pacholczyk 1992) :

$$\vee(\tau_i, \tau_j) = \max(\tau_i, \tau_j)$$
$$\wedge(\tau_i, \tau_j) = \min(\tau_i, \tau_j),$$
$$\sim\tau_j = \tau_{M-j}.$$

Let us choose Luckasiewicz material implication $\rightarrow_L$, so as to define a neighboring relation between two words of our knowledge base.

$$\tau_i \rightarrow_L \tau_j = \begin{cases} \tau_M & \text{if } i \leq j \\ \tau_{M-(i-j)} & \text{if } i > j \end{cases}$$

Example : M=7

L_7 = {not at all, very little, little, ∅, rather, very, totally}
= $\{\tau_1,\tau_2,\dots.\tau_7\}$

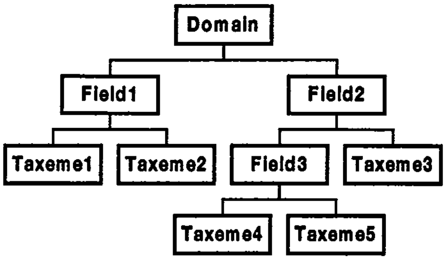
Then we obtain the following Luckasiewicz implication table (lines corresponding to τ_i, and columns to τ_j).

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
τ_1	τ_7	τ_7	τ_7	τ_7	τ_7	τ_7	τ_7
τ_2	τ_6	τ_7	τ_7	τ_7	τ_7	τ_7	τ_7
τ_3	τ_5	τ_6	τ_7	τ_7	τ_7	τ_7	τ_7
τ_4	τ_4	τ_5	τ_6	τ_7	τ_7	τ_7	τ_7
τ_5	τ_3	τ_4	τ_5	τ_6	τ_7	τ_7	τ_7
τ_6	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_7
τ_7	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7

with: τ_1 : not at all true (totally false, Boolean false)
τ_2 : very little true (very false)
τ_3 : little true (rather false)
τ_4 : ∅ true (∅ false)
τ_5 : rather true (little false)
τ_6 : very true (very little false)
τ_7 : totally true (not at all false, Boolean true)

Accounting for the two notions of point of view and graduality

The starting corpus is formed by all words in the current synonyms dictionary. We use the categorization in "Domain-Fields-Taxemes" D-F-T proposed by F. Rastier (Rastier 1991), (Rastier, Cavazza and Abeillé 1994).



The most general class is the *domain*. Each *domain* is related to a well-defined social activity (e.g. anatomy, chemistry, cooking ...). About 300 to 400 *domains* can be found in written languages of developed countries (Rastier 1991), (Rastier, Cavazza and Abeillé 1994).

The minimal class is the *taxeme*, where both specific sema and the least generic seme are defined.

The *field* is a structured set of *taxemes*. To each differential seme corresponds a *field*.

The interest of such a decomposition is that inside one *domain*, there is no lexical polysemy, because the polysemy results from the multiplicity of *domains*.

The tree is constructed so as to further reduce the class related to the *domain*, until the class obtained in a *taxeme* is only composed of sememes which, in some contexts (specified by the *taxeme*) can be considered as equivalent.

Once the *domains* have been listed, the system stores the words of the corpus in the *domain*(s) in which they appear.

Each word is stored with its definition as sema, described in an appropriate formalism (Redjai 1995).

The system asks the expert to specify the differential sema which are necessary to determine the different *fields* of the tree. Then, it builds a D-F-T tree and presents it to the expert for validation.

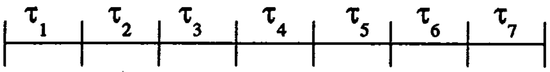
Once the tree has been validated, the words belonging to one *taxeme* are equivalent, and may actually be interchangeable.

At this stage, if one wants to find all the synonyms of a word W_i (in a precise *domain*), all the words belonging to the *taxeme* where W_i appears are candidates. Yet this list may be rather large and may contain words that are only remote synonyms of W_i.

The solution we propose consists in retaining only the best synonyms, those which come closest to W_i. Here appears a notion of proximity. But this proximity is in fact relevant only with respect to a particular *point of view*. The intervention of the expert is necessary at this level, so as to provide the system with the different *points of view* under which the words of the *taxeme* may be considered.

In fact, these *points of view* correspond to generic properties of the *taxeme*. Furthermore, the expert is able to determine which word in the *taxeme* verifies best such or such property (*point of view*), and which verifies it least.

He is also able to place all the words of that list on a scale (τ_1: not at all, . . . , τ_7: totally) :



The evaluation given by the expert is interpreted as follows :

W_i verifies the property p_j to a degree α
means that $W_i \in_\alpha p_j \Rightarrow W_i$ is p_j is τ_α-true.

We need the following notions between truth-degrees to define the neighboring degree of two sememes, and the distance between them :

- **Neighboring**

  $N(\tau_i, \tau_j) - \mathrm{Min}\ \{\tau_i \rightarrow_L \tau_j, \tau_j \rightarrow_L \tau i)$
  Two equal degrees have neighboring degree τ_M

- **Distance**

  $D(\tau_i, \tau_j) - {\sim}N(\tau_i, \tau_j)$
  Two equal degrees have distance τ_0

- **τ_γ-neighboring**

  The elements y such that x $\aleph_\gamma$ y
  constitute the τ_γ-neighboring of x
  It is the set of y that have at the most
  distance ${\sim}\tau_\gamma$ from x.

We can now, in the same manner, define inside a *taxeme* t_x, the neighboring degree N_w of two sememes, as well as the distance D_w between them with respect to a *point of view* p_x (Redjai 1997) :

$$N_W(S_i,S_j)_{px} = N(\tau_{\alpha i}, \tau_{\alpha j})$$

$$D_w(S_i,S_j)_{px} = D(\tau_{\alpha i}, \tau_{\alpha j})$$

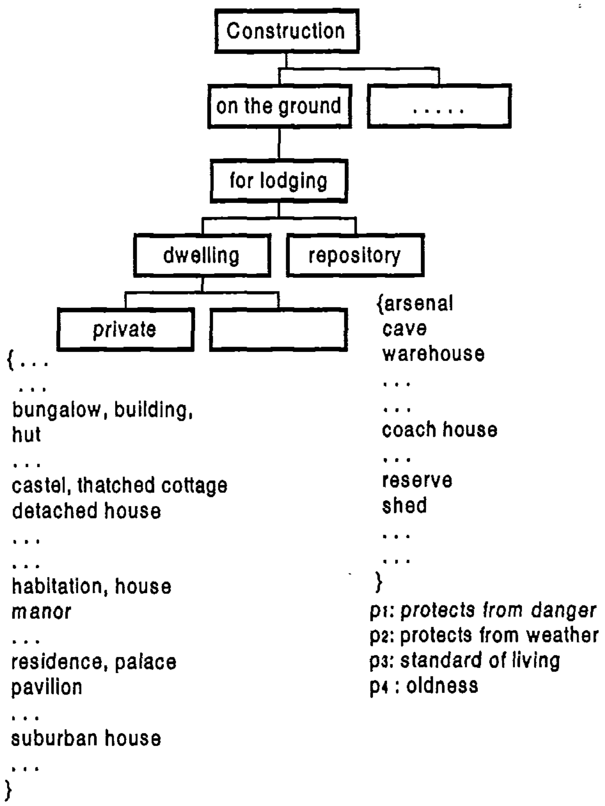
with: "S_i is p_x" is $\tau_{\alpha i}$-true

and: "S_j is p_x" is $\tau_{\alpha j}$-true

The sememes W_j such as $W_i \mathbb{N}^w W_j$ constitute the τ_γ-neighboring of W_i. It is the set of W_j which have at the most distance $\sim\tau_\gamma$ from W_i.

$$
\boxed{
\begin{array}{l}
\text{Synonyms}(Wi, p_x) = \tau_\gamma\text{-neighboring of } W_i \\
\quad = \{\ Wj \quad \text{such as} \quad \text{“}W_i \text{ is } p_x\text{” is } \tau_{\alpha i}\text{-true} \\
\qquad\quad \text{and} \quad \text{“}W_j \text{ is } p_x\text{” is } \tau_{\alpha j}\text{-true} \\
\qquad\quad \text{and} \quad \tau_{\alpha j} \in \tau_\gamma\text{-neighboring of } \tau_{\alpha i} \\
\qquad \} \\
\gamma = 7, 6, 5, 4
\end{array}
}
$$

Example



With the following propositions :

.....
“bungalow is p_1” is τ_3-true
“building is p_1 is τ_4-true
“hut is p_1 is τ_3-true
“castle is p_1” is τ_6-true
“detached house is p_1” is τ_4-true
“house is p_1” is τ_6-true
“manor is p_1” is τ_6-true
“residence is p_1” is τ_5-true
“palace is p_1” is τ_6-true
“pavilion is p_1” is τ_4-true
“house is p_1” is τ_6-true
.....

Since **“residence is p_1”** is τ_5-true,

$$
\begin{array}{lll}
\text{Synonyms}(\textbf{residence}, p_1) = \{ & & W_j \text{ such as} \\
& & \text{“}W_j \text{ is } p_1\text{” is } \tau_{\alpha_j}\text{-true} \\
& \text{with} & \tau_{\alpha_j} \in \tau_\gamma\text{-neighboring of } \tau_5 \\
& \text{and} & \gamma = 5 \\
& \} &
\end{array}
$$

$$\tau_5 \aleph_5 \tau_{\alpha_j} \Rightarrow \{\tau_5, \tau_4, \tau_6, \tau_3, \tau_7\}$$

We first propose all the words Wj such as :

“Wj is p_1” is τ_5-true

and then, those which are

“Wj is p_1” is τ_4-true,

and finally those which are

“Wj is p_1” is τ_6-true, or τ_3-true.or τ_7-true

So :
Synonyms(**residence**, p_1) = {building, detached house, pavilion, castle, house, manor, ...}

Conclusion

The proposed system brings a genuine solution for the problem which we currently meet in the synonyms dictionaries. In fact, the system achieves in proposing in a limited period of time the wanted synonym. Furthermore, the use of the notion of *point of view* enriches the propositions given by the system.

The tests in progress (two domains: construction - means of transport) give rather satisfying results. They should be performed for a dozen domains in order to evaluate the reliability of this system before extending it to all the *domains* of the language.

References

Akdag, H. 1992. Une approche logique du raisonnement incertain. Thèse de doctorat d'état, université Pierre et Marie Curie, Paris VI.

Akdag, H., De Glas, M. and Pacholczyk D. 1992. Qualitative theory of uncertainty. Fundamenta Informaticae 17: 333-347

Coseriu, E. 1976. L'étude fonctionnelle du vocabulaire. Cahiers de lexicologie 27:30-51.

De Glas, M. 1984. Representation of Luckasiewicz many-valued algebras. The atomic case Fuzzy Sets and Systems 14.

Dubois, D. 1991. Sémantique et cognition. Editions du CNRS, Paris.

Katz, J. 1972. Semantic Theory. New York, Harper & Row.

Katz, J. and Fodor, J.A. 1963. Structure of a semantic theory, Language 38:170-210.

Pacholczyk, D. 1992. Contribution au traitement logico-symbolique de la connaissance. Thèse de doctorat d'état, université Pierre et Marie Curie, Paris VI.

Rastier, F. 1991. Sémantique et recherches cognitives. PUF, Paris.

Rastier, F., Cavazza, M. and Abeillé, A. 1994. Sémantique pour l'analyse. De la linguistique à l'informatique, Masson, Paris.

Redjai, F.1995. Modélisation de la synonymie - Rapport interne, LAFORIA.

Redjai, F. 1997. Prise en compte de la gradualité dans la recherche automatique des synonymes, Rapport interne, LAFORIA.

Pottier, B. 1962. Sémantique des éléments de relations, Klincksieck, Paris.

Pottier, B. 1974. Linguistique générale. Théorie et description, Klincksieck, Paris.