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), (Cosneriu 1976). It adopts an onomasiological method. The sememes are analyzed according to their differences inside a minimal semantic class (the taxeme), constructed 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 De Morgan lattice Ln = {τ1, τ2, ..., τn} of truth-degrees (τi ≤ τj iff i ≤ j) going from τ1 (“not at all”) to τn (“totally”), supplied with the ∨, ∧, and ¬ operators (Akdag 1992), (Akdag, De Glas and Pacholczyk 1992):

\[ \begin{align*} 
\land (\tau_i, \tau_j) &= \min(\tau_i, \tau_j), \\
\lor (\tau_i, \tau_j) &= \max(\tau_i, \tau_j), \\
\neg \tau_i &= \tau_{n-i+1}. 
\end{align*} \]

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

\[ \tau_i \rightarrow_{\land} \tau_j = \begin{cases} 
\tau_n & \text{if } i \leq j \\
\tau_{n+\delta} & \text{if } i > j 
\end{cases} \]

Example : M=7

\[ L = \{ \text{not at all, very little, little, \emptyset, rather, very, totally} \} = \{ \tau_1, \tau_2, ..., \tau_7 \} \]

Then we obtain the following Luckasiewicz implication table (lines corresponding to \( \tau_i \) and columns to \( \tau_j \)).

<table>
<thead>
<tr>
<th></th>
<th>( \tau_1 )</th>
<th>( \tau_2 )</th>
<th>( \tau_3 )</th>
<th>( \tau_4 )</th>
<th>( \tau_5 )</th>
<th>( \tau_6 )</th>
<th>( \tau_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau_1 )</td>
<td>( \tau_1 )</td>
<td>( \tau_2 )</td>
<td>( \tau_3 )</td>
<td>( \tau_4 )</td>
<td>( \tau_5 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_2 )</td>
<td>( \tau_2 )</td>
<td>( \tau_2 )</td>
<td>( \tau_3 )</td>
<td>( \tau_4 )</td>
<td>( \tau_5 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_3 )</td>
<td>( \tau_3 )</td>
<td>( \tau_3 )</td>
<td>( \tau_3 )</td>
<td>( \tau_4 )</td>
<td>( \tau_5 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_4 )</td>
<td>( \tau_4 )</td>
<td>( \tau_4 )</td>
<td>( \tau_4 )</td>
<td>( \tau_4 )</td>
<td>( \tau_5 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_5 )</td>
<td>( \tau_5 )</td>
<td>( \tau_5 )</td>
<td>( \tau_5 )</td>
<td>( \tau_5 )</td>
<td>( \tau_5 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_6 )</td>
<td>( \tau_7 )</td>
</tr>
<tr>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
<td>( \tau_7 )</td>
</tr>
</tbody>
</table>

with: \( \tau_1 \) : not at all true (totally false, Boolean false)
\( \tau_2 \) : very little true (very false)
\( \tau_3 \) : little true (rather false)
\( \tau_4 \) : \( \emptyset \) true (\( \emptyset \) false)
\( \tau_5 \) : rather true (little false)
\( \tau_6 \) : very true (very little false)
\( \tau_7 \) : 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 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. 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 \( (\tau_1: \text{not at all}, \ldots, \tau_7: \text{totally}) \):

\[
\begin{array}{cccccccc}
\tau_1 & \tau_2 & \tau_3 & \tau_4 & \tau_5 & \tau_6 & \tau_7 \\
\end{array}
\]

The evaluation given by the expert is interpreted as follows:

\( W_i \) verifies the property \( p_j \) to a degree \( \alpha \) means that \( W_i \in_\alpha p_j \Rightarrow W_i \) is \( p_j \) true.

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

- **Neighboring**

\[ N(\tau_i, \tau_j) = \min \{ \tau_i - \tau_j, \tau_j - \tau_i \} \]

Two equal degrees have neighboring degree \( \tau_0 \).

- **Distance**

\[ D(\tau_i, \tau_j) = -N(\tau_i, \tau_j) \]

Two equal degrees have distance \( \tau_0 \).

- **\( \tau \)-neighboring**

The elements \( y \) such that \( x \sim y \) constitute the \( \tau \)-neighboring of \( x \). It is the set of \( y \) that have at the most distance \( \tau \) from \( x \).

We can now, in the same manner, define inside a taxeme \( t \), 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):

\[
N_w(S_i, S_j)_{p_x} = N(\tau_{ai}, \tau_{aj})
\]

\[
D_w(S_i, S_j)_{p_x} = D(\tau_{ai}, \tau_{aj})
\]

with: "\( S_i \) is \( p_x \)" is \( \tau_{ai} \)-true

and: "\( S_j \) is \( p_x \)" is \( \tau_{aj} \)-true

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 sema are defined.

The *field* is a structured set of *taxemes*. To each differential sema 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 sememes $W_i$ such as $W_i \in N^w W_i$ constitute the $\tau$-neighboring of $W_i$. It is the set of $W_j$ which have at the most distance $\sim \tau$ from $W_i$.

\[
\text{Synonyms}(W_i, p_a) = \{ W_j \text{ such as } \text{"}W_i \text{ is } p_a\text{" is } \tau_a\text{-true} \\
\text{and } \text{"}W_j \text{ is } p_a\text{" is } \tau_a\text{-true} \\
\text{and } \tau_{ai} \in \tau\text{-neighboring of } \tau_{ai} \}
\]

$\gamma = 7, 6, 5, 4$

**Example**

![Diagram]

<table>
<thead>
<tr>
<th>Construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>on the ground</td>
</tr>
<tr>
<td>for lodging</td>
</tr>
<tr>
<td>dwelling</td>
</tr>
<tr>
<td>repository</td>
</tr>
<tr>
<td>private</td>
</tr>
<tr>
<td>arsenal</td>
</tr>
<tr>
<td>cave</td>
</tr>
<tr>
<td>warehouse</td>
</tr>
<tr>
<td>....</td>
</tr>
<tr>
<td>bungalow, building, hut</td>
</tr>
<tr>
<td>castel, thatched cottage</td>
</tr>
<tr>
<td>detached house</td>
</tr>
<tr>
<td>....</td>
</tr>
<tr>
<td>habitation, house</td>
</tr>
<tr>
<td>manor</td>
</tr>
<tr>
<td>....</td>
</tr>
<tr>
<td>residence, palace</td>
</tr>
<tr>
<td>pavilion</td>
</tr>
<tr>
<td>....</td>
</tr>
<tr>
<td>suburban house</td>
</tr>
</tbody>
</table>

With the following propositions:

- "bungalow is $p_1$" is $\tau_3$-true
- "building is $p_1$" is $\tau_4$-true
- "hut is $p_1$" is $\tau_5$-true
- "castle is $p_1$" is $\tau_6$-true
- "detached house is $p_1$" is $\tau_7$-true
- "house is $p_1$" is $\tau_8$-true
- "manor is $p_1$" is $\tau_9$-true
- "residence is $p_1$" is $\tau_{10}$-true
- "palace is $p_1$" is $\tau_{11}$-true
- "pavilion is $p_1$" is $\tau_{12}$-true
- "house is $p_1$" is $\tau_{13}$-true

Since "residence is $p_1$" is $\tau_{13}$-true,

\[
\text{Synonyms(residence, } p_1) = \{ W_j \text{ such as } \\
\text{"}W_i \text{ is } p_1\text{" is } \tau_a\text{-true} \\
\text{with } \tau_{ai} \in \tau\text{-neighboring of } \tau, \\
\text{and } \gamma = 5 \}
\]

$\tau_5 \& \tau_{13} \Rightarrow \{ \tau_5, \tau_6, \tau_8, \tau_9, \tau_7 \}$

We first propose all the words $W_j$ such as:

- "Wj is $p_1$" is $\tau_5$-true

and then, those which are

- "Wj is $p_1$" is $\tau_6$-true,

and finally those which are

- "Wj is $p_1$" is $\tau_8$-true, or $\tau_5$-true, or $\tau_6$-true

So:

\[
\text{Synonyms(residence, } p_1) = \{ \text{building, detached house, pavilion, castle, house, manor} \ldots \}
\]

**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


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.
