Formal versus Distributional Semantics of Natural Language: a Survey

ANCA DINU

Abstract - This paper discusses the relation of the two major research streams in natural language semantics: Formal Semantics and Distributional Semantics. At a first glance, their relation is apparently contrastive, if not even in plain opposition. At a closer look, however, it turns out that the two frameworks have in fact a complementarity relation. Both theories, and natural language semantics in general, can benefit from integrating the two, using each other’s specific methods and ideas. For instance, Formal semantics could integrate distributional information for representing content words, while Distributional Semantics could employ key ideas about compositionality from Formal Semantics to represent phrases, sentences or discourse, and grammatical words. Few efforts were directed towards integrating lexical distributional information into Formal Semantics, in sharp contrast to the concentrated and abundant efforts made for obtaining a compositional model in Distributional Semantics. The paper points to the main literature representing these research directions.

Key words and phrases:
Formal semantics, distributional semantics, natural language, compositionality.

1 Introduction

This paper discusses the relation between the two mainstream research directions in Natural Language Semantics, namely Formal Semantics and Distributional Semantics. Thus, it ascribes itself to the greater effort currently being carried out on an international level of merging the two directions.

The most general goal of Natural language semantics is to provide a theory of meaning, that is a detailed specification of the knowledge that a native speaker has about his/her own language [17]. In doing this, a theory of meaning has to provide a way to assign meaning to all the different words in the language and a mechanism by means of which all these meanings can be combined into larger expressions to form the meaning of phrases, sentences or discourses.

The dual nature of natural language meaning has long been noticed, linguistic signs having both a reference and a sense [21], which are built compositionally. For instance, (1) has the same reference and the same sense, while (2) has the same reference and different sense.
(1) Mark Twain is Mark Twain.
(2) Mark Twain is Samuel Clemens.

The origin of the two research directions of natural language semantics is precisely this double load of meaning: Formal Semantics studies meaning as reference, while Distributional Semantics focuses on meaning as sense, adopting the so called language as use view.

In the following, we will discuss separately Formal Semantics and Distributional Semantics in section 2 and 3, respectively. In section 4, we will present the main research effort towards obtaining a compositional distributional semantic model. Finally, we will draw conclusions in section 5.

2 Formal Semantics

Up until the last decade, the most successful semantic framework for natural language was formal (model-theoretic) semantics. In the last 40 years, Formal Semantics has offered an elegant and elaborate framework for formalizing those linguistic phenomena which occur systematically and productively. It thus contributed decisively to understanding the functioning of the natural language. It also paved the way for applications in computational semantics, which comprise techniques of automatic construction for the representations of human language expressions, in order to be able to further make inferences. Thus, notorious applications of formal and computational semantics include model checking and theorem proving.

Most notably, Formal Semantics has successfully treated those linguistic phenomena related to the compositionality of natural languages. The famous fregean principle of compositionality occupies a central place in the programme of Formal Semantics and stipulates that the meaning of a sentence is determined by the meaning of the words it contains and the way in which they combine (i.e. through syntax). Syntax is needed because the linear order of the words in a sentence hide the role that different kinds of words play in the building of the meaning of the whole.

Montague was the first to formalize the semantics of natural language in its famous *English as a formal language* [39] and *The proper treatment of quantification in ordinary English* [40]. He shows how to compositionally construct representations for the meaning of natural language expressions, by specifying a semantic rule for each syntactic rule. He stipulates that to understand the meaning of a sentence, in a model-theoretic way, one should be able to determine the conditions in which it is true, and thus to be able to test them in a given model. For instance, for a typical example of a quantified sentence like:

(3) Every man loves Mary.
the truth-conditions are: for each $x$, if $x$ is in the set of men and $m$ represents the denotation of (contextually unique or salient) individual Mary, then the pair $<x, m>$ is in the relation *loves*, which is formally written as:

$$\forall x. (\text{man}(x) \Rightarrow \text{loves}(x, m))$$

Formal Semantics typically uses Lambda Calculus as a means of combining meaning guided by the syntactic operations. For instance, the meaning of sentence (3) is obtained in a bottom up compositional manner, as in figure 1.

Montague’s first formal semantic theory was the source of inspiration for a whole programming of Formal Semantics, which included over the years theories such as Dynamic Montague Grammar [23], Dynamic Predicate Logic [24], Discourse Representation Theory [27], or Continuation Semantics [2],[15], [14]. All these theories focus mostly on semantic phenomena such as negation, quantification, entailment, determiners, prepositions, pronouns, anaphora, coordination or auxiliary verbs.

Nevertheless, the programme of Formal Semantics has had nothing to say about the meaning of words which have their own semantic content, such as nouns, adjectives, verbs, and adverbs, words which are traditionally dealt with by lexical semantics. Any formal semantic theory assumes from the start that these content words are already disambiguated, which is by no means a simple task (see the following famous example for the adjective *brown*). Since many semantic problems appear from the lexicon, i.e. from the interpretation of (ambiguous) lexical words and not from the interpretation of function words, a new semantic theory was needed to fill this void.

It is worth noting that important attempts to model the meaning of a lexical word have been made within some Formal Semantics theories, two of which being The Generative Lexicon Theory [42] and, more recently, the Segmented Discourse Representation [1], who employed the elegant and sophisticated concept of continuations together with a mechanism of type shifting.

Still, in spite of these efforts, the main problem of lexical semantics persists to a great extent: even when considering the plentiful sub-regularities found in the lexicon for lexical words, a manual analysis is simply not feasible [6]. For instance, each adjective acts on nouns in a different way. Lahav [28] exemplifies this with his famous example of the adjective *brown*:

In order for a cow to be brown most of its body’s surface should be brown, though not its udders, eyes, or internal organs. A brown crystal, on the other hand, needs to be brown both inside and outside. A book is brown if its cover, but not necessarily its inner pages, are mostly brown, while a newspaper is brown
only if all its pages are brown. For a potato to be brown it needs to be brown only outside. . . Furthermore, in order for a cow or a bird to be brown the brown color should be the animal’s natural color, since it is regarded as being ‘really’ brown even if it is painted white all over. A table, on the other hand, is brown even if it is only painted brown and its ‘natural’ color underneath the paint is, say, yellow. But while a table or a bird are not brown if covered with brown sugar, a cookie is. In short, what is to be brown is different for different types of objects. To be sure, brown objects do have something in common: a salient part that is wholly brownish. But this hardly suffices for an object to count as brown. A significant component of the applicability condition of the predicate ‘brown’ varies from one linguistic context to another.

What happens with brown is replicated by the large majority of adjective-noun combinations. Treating them all like ‘idioms’ would mean to turn the exception into the rule. No formal theory until now has been able to surpass this problem, that makes building (either automatically, or by hand) a fine-grained lexicon intractable.

3 Distributional Semantics

The lack of proper treatment of lexical words semantics lead to the development of the more recent Distributional Semantics, which belongs to the more general Vector Space Model Semantics class (see [47] for an excellent survey of Vector Space Models).

While there are other related frameworks for representing lexical infor-
mation, like network semantics [10] (where words are represented as nodes in a graph interconnected by semantic relationships between concepts, and where word meaning is expressed by the number and type of connections to other words) and feature-based semantics [46] (where a word is represented by a distribution of numerical values over the feature set), Distributional Semantics is the onely one that can be fully automatically obtained at a large scale from corpora.

Distributional Semantics manages to successfully represent lexical words and, in general, lexical semantic phenomena, starting from the idea that the statistics with which words occur in large corpora determines, to a great extent, the meaning of those words. This idea is founded at least as early as Firtf (1957) [20] "you shall know a word by the company it keeps" and Harris (1954) [26] "words that occur in similar contexts tend to have similar meanings". Thus, in Distributional Semantics, the meaning of a word is given by the context in which it appears and, in consequence, two or more words are said to be similar in meaning if they appear in similar contexts. This observation is named "the distributional hypothesis" and stipulates that semantic similarity means distributional similarity.

Practically, the semantic similarity of two words $w_1$ and $w_2$ is calculated as the co-occurrence count of these words in a large corpus G. The co-occurrence is the number of times that $w_1$ and $w_2$ occur: in a linguistic relationship with each other (e.g., $w_1$ is a modifier of $w_2$), or in the same sentence, or in the same document, or within a distance of at most $k$ words of each other.

In a world in which access to huge amounts of data is no longer a problem, corpora become available for experimenting within Distributional Semantics. One of the most popular (and freely available) is the Wacky corpus [3], which contains over 2 billion words, is lemmatized and POS-tagged. For instance, given this corpus and the co-occurrence defined as occurrence within $k = 10$ words of each other, Schütze [45] obtains the values in table 3.

Visually, one can represent this like in figure 2. In this example, rich and poor act as axes of the space, which are usually called dimension words. Words like silver, gold, disease and society are represented as vectors in the space formed by the two axes. In Distributional Semantics, the number of dimension words may vary from hundreds to thousands, thus obtaining a high-dimensional space. Dimension words are usually taken to be the most

<table>
<thead>
<tr>
<th>co-occurrence</th>
<th>silver</th>
<th>gold</th>
<th>disease</th>
<th>society</th>
</tr>
</thead>
<tbody>
<tr>
<td>rich</td>
<td>186</td>
<td>170</td>
<td>17</td>
<td>143</td>
</tr>
<tr>
<td>poor</td>
<td>34</td>
<td>59</td>
<td>162</td>
<td>228</td>
</tr>
</tbody>
</table>

Table 1: Co-occurrence of some words in English Wikipedia
frequent content words in a corpus (functional or grammatical words are excluded). In this word space, the similarity between two words is usually measured with the cosine of the angle between them: the smaller the angle, the more similar the words. In figure 2, silver and gold are similar, because of the small angle between them; silver and society are not so similar, having a bigger angle between them; silver and disease are even less similar. In this manner, given a large corpus, one can represent as a vector in a word space any word that has its own semantic content. One can also compute its nearest neighbors.

In a similar manner, one can represent in a vector space not only words, but other entities like phrases, sentences, paragraphs, documents, even entire books.

Moreover, Distributional Semantics presents a number of advantages:

- it can be obtained completely automatically;
- the required representations are simple (it only needs a corpus and memory space for storing the vectors);
- it is independent of the language it is applied to;
- it is cognitively plausible.

Also, Distributional Semantics has an impressive number of applications, among which we mention document retrieval and classification, question answering, automated thesaurus construction, machine translation [16], [47], semantic priming [29], [34], discourse comprehension [29], judgments of essay quality [30], predicting similarity judgments [22], [29], [37] and association [22], categorizing nominal concepts into hypernyms, generating salient properties of concepts (and qualia of nouns), capturing intuitions about the thematic fit of verb arguments and even spotting the alternation classes of verbs [5], [29], [32], [34], [41], a.o.

As it was expected, however, Distributional Semantics also has its disadvantages, such as:

- it is not a generative model;
- it requires a lot of ad-hoc parameters;
- the meaning of homonym words such as bank is calculated only as a mean of its distinct (unrelated) meanings;
- non-specific words (words that occur in a large variety of different contexts and have few or no specific semantic associations), such as person cannot be fairly represented;
Figure 2: Representing words in space

- it does not realistically represent grammatical and functional words (e.g. negation, quantifiers, words with anaphoric or variable meaning like pronouns and proper names, etc.).

But maybe the biggest disadvantage is the fact that in Distributional Semantics the problem of compositionality has yet to be reasonably resolved: the Distributional Semantics models create representations of words or groups of words as vectors which contain their patterns of co-occurrence in large corpora, but they do not present any obvious way of treating combinations of words which appear sparsely or even do not appear at all in corpora [48].

Moreover, in Distributional Semantics, words contribute equally to the meaning of the expression they form, regardless of the way in which they combine. Combining words into phrases by the obvious vector averaging (which is equivalent with addition under the cosine measure) is insensitive to word order and to syntax in general. This way, a sentence such as "The cat follows the mouse" has the same meaning as the sentence formed out of the same words but with a different ordering of the words, such as "The mouse follows the cat". In the following section we summarize the numerous attempts to overcome this problem.

4 Compositional Distributional Models

Attempts to find a way of incorporating compositionality into Distributional Semantics (and, more generally, in Space Models), so that representations could also be generated for groups of words not yet observed in the corpus, currently form the object of a rich and ongoing research.
In one of the most influential papers about Compositional Semantic Models, Mitchell and Lapata [38] present a list of compositional models in Distributional Semantics, from the basic vector addition, to tensor product and convolution. Specifically, they formulate composition as a function of two vectors \( u \) and \( v \) and their syntactic relation \( R \):
\[
p = f(u, v, R)
\]

While composing vectors as addition does not take into consideration any syntactic information, the multiplicative class of functions allows us to think of one representation as modifying the other. This idea is fundamental in Formal Semantics, where different syntactic structures are given different function types. Unfortunately, tensor products [8], [9] grow exponentially as more vectors are combined, making them computationally intractable and implausible as models of human concept combination. However, this problem can be bypassed (see [38] for a discussion).

First steps towards representing whole sentences or even discourse in Distributional Semantics comprise representing noun-noun compounds and adjectival phrases. The representation and similarity of adjectival phrases in Distributional Semantics has been investigated in numerous works (such as [4], [18], [25], [31]). It has multiple applications, among which is disambiguation of syntactic structure [48]. For instance, there are two syntactic structures that can be associated to an expression such as "live fish transporter", e.g. ((live fish) transporter) and (live (fish transporter)). Stanford Parser [44], which lacks semantic information, incorrectly predicts the second structure. Obviously, the semantic plausibility of the two parsing versions should be considered.

In the same line of research, Manning [35] introduces recursive neural networks which learn compositional vector representations for expressions and sentences; Dinu and Ciobanu [7] propose an alternative method of measuring semantic similarity between words; recently, Dinu [12], Dinu and Baroni [13] and Li et al. [33] have also proposed distributional models for the generation of expressions. In an excellent recent paper, Baroni et al. [6] present the results of the Programm for Compositional Distributional Semantics.

The attention given to the problem of incorporating compositionality into Distributional Semantics is also demonstrated by the introduction, for the first time, of a track at the SemEval-2014 workshop dedicated to the Evaluation of Compositional Models in Distributional Semantics on phrase similarity and entailment (see [36]).

Unfortunately, a similar research effort for including lexical information of content words in the framework of Formal Semantics is lacking in the literature. One of the most important attempts to incorporate fine-grained lexical information is Pustejowsky’s Generative Lexicon [42]. Although it is definitely a first elegant step towards formally representing lexical infor-
nation, an important disadvantage of this semantic framework is its lack of feasibility, needing a huge human qualified annotation effort. Even supposing that such an effort is possible, the subjective nature of personal decision for annotating fine-grained sub-regularities hinders on the success of such an enterprise. Also, lexical aspects of combining phrases, such as noun-noun or adjective-noun (see for instance [11]) have been studied in a Formal Semantic framework, but these efforts are no match for the concentrated effort to include compositionality in Distributional Semantic models.

Moreover, semantics (of any type) is still far from providing a proper understanding of the interpretation of natural language. Consequently, computers still understand very little of natural language semantics, thus a semantic representation of an arbitrary piece of text cannot be automatically generated, given its syntactic structure.

For syntax, the decisive step in assigning syntactic structures to arbitrary sentences, with a reasonable accuracy, was the emergence of the probabilistic parsers. Although distribution driven models, such as Distributional Semantics, create the conditions for a similar step in semantics, so far, the problems of compositionality and of handling functional words still prevents decisive progress in this direction.

5 Conclusions

The increased international research effort of bringing together Formal and Distributional Semantics focuses on integrating their advantages and emphasizes their complementary character. Both semantic theories can benefit from mutual use of each theory’s specific methods. Ideally, the formal side could better incorporate lexical and distributional information in the case of words which have their own semantic content like nouns, verbs, adjectives or adverbs, while the distributional side could borrow models from the formal side to better treat compositionality and function words like negation, quantifiers, pronouns, etc.

6 Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS - UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

References


Anca Dinu
University of Bucharest, Faculty of Foreign Languages and Literatures
Str. Edgar Quinet nr. 5-7, cod 010017, sector 1, București
E-mail: anca.dinu@lls.unibuc.ro; anca_d_dinu@yahoo.com