Semantic Domains and Linguistic Theory

Alfio Gliozzo
University of Rome Tor Vergata
Department of Computer Science, System and Production
00133 Roma (Italy)
gliozzo@itc.it

Abstract

This paper is about the relations between the concept of Semantic Domain and the "Theory of Semantic Fields", a structural model for lexical semantics proposed by Jost Trier at the beginning of the last century. The main limitation of the Trier's notion is that it does not provide an objective criterion to aggregate words around fields, making the overall model too vague, and then unuseful for computational purposes. The notion of Semantic Domain improves that of Semantic Field by providing such a criterion. In particular, the structuralist approach in semantics has been connected to the Wittgenstein's *meaning-is-use* assumption, providing an objective criterion to infer Semantic Domains from corpora relying on a lexical coherence assumption. The task based evaluation we did for our claims shows that the notion of Semantic Domains is effective because it allows to define an uniform methodology to deal with many different Natural Language Processing tasks. In the paper we discuss the epistemological issues concerning the possibility of adopting a task based methodology to support linguistic theory, showing the case study of Semantic Domains in Computational Linguistics as a paradigmatic example for our claims.

1 Introduction

The predominant view in lexical semantic is the Saussure's structural semantics (de Saussure, 1922), claiming that a word meaning is determined by the "horizontal" paradigmatic and the "vertical" syntagmatic relations between that word and others in the whole language (Lyons, 1977). Structural assumptions are also widely adopted in Computational Linguistic. For example, many machine readable dictionaries describe the word senses by means of semantic networks representing relations among terms (e.g. WORDNET (Miller, 1990)). The main limitation of the "radical" structuralist view is that it is almost impossible to describe the associations among all the possible terms in a natural language, because the huge number of concepts and semantic relations among them.

The Semantic Fields Theory (Trier, 1931) goes a step further in the structural approach to lexical semantics by introducing an additional aggregation level and by delimiting to which extend paradigmatic relations hold. The basic assumption of this theory is that the lexicon is structured into Semantic Fields: semantic relations among concepts belonging to the same field are very dense, while concepts belonging to different fields are typically unrelated. In fact, a word meaning is established only by the network of relations among the terms of its field. Another property of great interest is that there exists a strong correspondence among Semantic Fields of different languages, while such a strong correspondence cannot be established among the terms themselves.

It has been observed that the main limitation of the Trier's notion is that it does not provide an objective criterion to aggregate words around fields, making the overall model too vague, and then unuseful for computational purposes. The notion of Semantic Domain improves that of Semantic Field by providing such a criterion. In particular, the structuralist approach in semantics has been connected to the *meaning-is-use* assumption introduced by Ludwig Wittgenstein in his celebrated "Philosophical Investigations" (Wittgenstein, 1965). A word meaning is its use into the concrete "form of life" where it is adopted, i.e. the *linguistic game*, in the Wittgenstein's terminology. Frequently co-occurring words in texts are then associated to the same linguistic game. It follows that fields can be identified from a corpus based analysis of the lexicon, exploiting the connections between linguistic games and Semantic Fields already depicted. The notion of Semantic Domain arises from this convergence, providing an objective criterion to identify semantically related words in texts, supported by a *lexical coherence* assumption, that we empirically corroborated in text in the earlier stages of our work.

The notion of Semantic Domain is intimately related to several phenomena in the language at both a lexical and a textual level. At a lexical level Semantic Domains can be used as a (shallow) model for lexical ambiguity and variability, while at a textual level semantic domains provide meaningful topic taxonomies that can be used to group texts into semantic clusters. In addition, the inherent multilingual nature of semantic domains allows an uniform representation of both the lexicon and the texts in most of the natural languages.

Exploiting Semantic Domains for Natural Language Processing (NLP) allowed us to improve sensibly the state-of-the-art in all those tasks in which they have been applied, providing and indirect evidence to support their linguistic properties. The major goal of this paper is to discuss the possibility of adopting such a task based methodology to support linguistic theory, showing the case study of Semantic Domains in computational linguistics as a successfully paradigmatic example of our methodology.

The paper is structured as follows. Section 2 is about the Semantic Fields Theory while Section 3 concerns the relations between this theory and the Wittgenstein's meaning-is-use assumption. Section 4 describes the concept of Semantic domains as the confluence of both perspectives, highlighting its technological impact in developing state-of-the-art systems for NLP, while Section 5 conclude the paper discussing the possibility of adopting the indirect task based evaluation to support linguistic theory.

2 The Theory of Semantic Fields

Semantic Domains are a matter of recent interest in Computational Linguistics (Magnini and Cavaglià, 2000; Magnini et al., 2002; Gliozzo et al., 2005a), even though their basic assumptions are inspired from a long standing research direction in structural linguistics started in the beginning of the last century and widely known as "The Theory of Semantic Fields" (Lyons, 1977). The notion of *Semantic Field* has proved its worth in a great volume of studies, and has been mainly put forward by Jost Trier (Trier, 1931), whose work is credited with having "opened a new phase in the history of semantics"(Ullmann, 1957).

In that work, it has been claimed that the lexicon is structured in clusters of very closely related concepts, lexicalized by sets of words. Word senses are determined and delimitated only by the meanings of other words in the same field. Such clusters of semantically related terms have been called Semantic Fields[1], and the theory explaining their properties is known as "The theory of Semantic Fields" (Vassilyev, 1974).

Semantic Fields are conceptual regions shared out amongst a number of words. Each field is viewed as a partial region of the whole expanse of ideas that is covered by the vocabulary of a language. Such areas are referred to by groups of semantically related words, i.e. the Semantic Fields. Internally to each field, a word meaning is determined by the network of relations established with other words.

There exists a strong correspondence among Semantic Fields of different languages, while such a strong correspondence cannot be established among the terms themselves. For example, the field of COLORS is structured differently in different lan-

[1] There is no agreement on the terminology adopted by different authors. Trier uses the German term *wortfeld* (literally "word field" or "lexical field" in Lyons' terminology) to denote what we call here semantic field.

guages, and sometimes it is very difficult, if not impossible, to translate name of colors, even whether the chromatic spectrum perceived by people in different countries (i.e. the conceptual field) is the same. Some languages adopt many words to denote the chromatic range to which the English term `white` refers, distinguishing among different degrees of "whiteness" that have not a direct translation in English. Anyway, the chromatic range covered by the COLORS fields of different languages is evidently the same. The meaning of each term is defined in virtue of its oppositions with other terms of the same field. Different languages have different distinctions, but the field of COLORS itself is a constant among all the languages.

Another implication of the Semantic Fields Theory is that words belonging to different fields are basically unrelated. In fact, a word meaning is established only by the network of relations among the terms of its field. As far as paradigmatic relations are concerned, two words belonging to different fields are then un-related. This observation is crucial form a methodological point of view. The practical advantage of adopting the Semantic Field Theory in linguistics is that it allows a large scale structural analysis of the whole lexicon of a language, otherwise infeasible. In fact, restricting the attention to a particular field is a way to reduce the complexity of the overall task of finding relations among words in the whole lexicon, that is evidently quadratic in the number of words.

The main limitation of the Trier's theory is that it does not provide any objective criterion to identify and delimitate Semantic Fields in the language. The author himself admits "what symptoms, what characteristic features entitle the linguist to assume that in some place or other of the whole vocabulary there is a field? What are the linguistic considerations that guide the grasp with which he selects certain elements as belonging to a field, in order then to examine them as a field?" (Trier, 1934). The answer to this question is an issue opened by the Trier's work.

3 Semantic Fields and the meaning-is-use view

In the previous section we have pointed out that the main limitation of the Trier's theory is the gap of an objective criterion to characterize Semantic Fields. The solutions we have found in the literature (Weisgerber, 1939; Porzig, 1934; Coseriu, 1964) rely on very obscure notions, of scarse interest from a computational point of view. To overcome such a limitation, in this section we introduce the concept of Semantic Domain (see Section 4).

The notion of Semantic Domain improves that of Semantic Fields by connecting the structuralist approach in semantics to the *meaning-is-use* assumption introduced by Ludwig Wittgenstein in his celebrated "Philosophical Investigations" (Wittgenstein, 1965). A word meaning is its use into the concrete "form of life" where it is adopted, i.e. the *linguistic game*, in the Wittgenstein's terminology. Words are then meaningful only if they are expressed into concrete and situated linguistic games that provide the conditions for determining the meaning of natural language expressions. To illustrate this concept, Wittgenstein provided a clarifying example describing a very basic linguistic game: "...Let us imagine a language ...The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones; there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words *block*, *pillar*, *slab*, *beam*. A calls them out; – B brings the stone which he has learnt to bring at such-and-such a call. – Conceive of this as a complete primitive language." (Wittgenstein, 1965)

We observe that the notions of linguistic game and Semantic Field show many interesting connections. They approach the same problem from two different points of view, getting to a similar conclusion. According to Trier's view, words are meaningful when they belong to a specific Semantic Field, and their meaning is determined by the structure of the lexicon in the field. According to Wittgenstein's view, words are meaningful when there exists a linguistic game in which they can be formulated, and their meaning is exactly their use. In both cases, meaning arises from the wider contexts in which words are located.

Words appearing frequently into the same linguistic game are likely to be located into the same field. In the previous example the words *block*, *pillar*, *slab* and *beam* have been used in a common linguistic game, while they clearly belong to the Semantic Field of BUILDING INDUSTRY. This example suggests that the notion of linguistic game provides a criterion to identify and to delimitate Semantic Fields. In particular, the recognition of the linguistic game in which words are typically formulated can be used as a criterion to identify classes of words composing lexical fields. The main problem of this assumption is that it is not clear how to distinguish linguistic games between each other. In fact, linguistic games are related by a complex network of similarities, but it is not possible to identify a set of discriminating features that allows us to univocally recognize them. "I can think of no better expression to characterize these similarities than 'family resemblances'; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way. - And I shall say: 'games' form a family" ((Wittgenstein, 1965), par. 67).

We observe that linguistic games are naturally reflected in texts, allowing us to detect them from a word distribution analysis on a large scale corpus. In fact, according to Wittgenstein's view, the content of any text is located into a specific linguistic game, otherwise the text itself would be meaningless. Texts can be perceived as open windows through which we can observe the connections among concepts in the real world. Frequently co-occurring words in texts are then associated to the same linguistic game.

It follows that the set of concepts belonging to a particular field can be identified from a corpus based analysis of the lexicon, exploiting the connections between linguistic games and Semantic Fields already depicted. For example, the two words *fork* and *glass* are evidently in the same field. A corpus based analysis shows that they frequently co-occur in texts, then they are also related to the same linguistic game. On the other hand, it is not clear what would be the relation among *water* and *algorithm*, if any. They are totally unrelated simply because the concrete situations (i.e. the linguistic games) in which they occur are in general distinct. It reflects on the fact that they are often expressed in different texts, then they belong to different fields.

Our proposal is then to merge the notion of linguistic game and that of Semantic Field, in order to provide an objective criterion to distinguish and delimitate fields from a corpus based analysis of lexical co-occurences in texts. We refer to this particular view on Semantic Fields by using the name Semantic Domains. The concept of Semantic Domain is the main topic of this work, and it will be illustrated more formally in the following section.

4 Semantic Domains

In our usage, Semantic Domains are common areas of human discussion, such as ECONOMICS, POLITICS, LAW, SCIENCE, which demonstrate lexical coherence. The Semantic Domain associated to a particular field is the set of domain specific terms belonging to it, and it is characterized by a set of *domain words* whose main property is to co-occur in texts.

An approximation to domains are Subject Field Codes, used in Lexicography (e.g. in (Procter, 1978)) to mark technical usages of words. Although this information is useful for sense discrimination, in dictionaries it is typically used only for a small portion of the lexicon. WORDNET DOMAINS (Magnini and Cavaglià, 2000) is an attempt to extend the coverage of domain labels within an already existing lexical database, WORDNET (Fellbaum, 1998). As a result WORDNET DOMAINS can be considered an extension of WORDNET in which synsets have been manually annotated with one or more domain labels, selected from a hierarchically organized set of about two hundred labels.

WORDNET DOMAINS represents the first attempt to provide an exhaustive systematization of the concept of Semantic Field and its connections to the textual interpretation depicted in section 3. It allowed us to start an empirical investigation about the connections between the textual and the lexical counterparts of Semantic Domains. First we concentrated on corroborating a lexical-coherence assumption, claiming that a great percentage of the concepts expressed in the same text belong to the same domain. Lexical coherence is then a basic property of most of the texts expressed in any natural language and it allows us to disambiguate words in context by associating domain specific senses to them. Otherwise stated, words taken out of context show domain polysemy, but, when they occur into real texts, their polysemy is solved by the relations among their senses and the domain specific concepts occurring in their contests.

Intuitively, texts may exhibit somewhat stronger or weaker orientation towards specific domains, but it seems less sensible to have a text that is not related to at least one domain. In other words, it is difficult to find a “generic” text. This intuition is largely supported by our data: all the texts in SemCor [2](Landes et al., 1998) exhibit concepts belonging to a small number of relevant domains, demonstrating the domain coherence of the lexical-concepts expressed in the same text. In particular, 34.5 % of nouns in co-occurring in the same texts in SemCor are annotated with the same domain label, while about 40% refer to generic concepts. The conclusion of this experiment is that there exists a strong tendency for the lexicon in texts to be aggregate around a specific domain. As we will see later in the paper, such a tendency should be presupposed to allow lexical disambiguation.

Then we investigated the relations between Semantic Domains and lexical ambiguity and variability, the two most basic and pervasive phenomena characterizing lexical semantics. The different senses of ambiguous words should be necessarily located into different domains, because they are characterized by different relations with different words. On the other hand, variability can be modeled by observing that synonymous terms refer to the same concepts, then they will necessarily belong to the same domain. Thus, the distribution of words among different domains is a relevant aspect to be taken into account to identify word senses. Understanding words in contexts is mainly the operation of locating them into the appropriate semantic fields.

To corroborate these assumptions we developed a Word Sense Disambiguation (WSD) procedure relying on domain information only, named Domain Driven Disambiguation (DDD) (Magnini et al., 2001; Gliozzo et al., 2004). The underlying hypothesis of the DDD approach is that information provided by domain labels offers a natural way to establish associations among word senses in a certain text fragment, which can be profitably used during the disambiguation process. DDD is performed by selecting the word sense whose Semantic Domain maximize the similarity with the domain of the context in which the word is located. For example, the word *virus* is ambiguous between its `Biology` and `Computer Science` senses, and can be disambiguated by assigning the correct domain to the contexts where it actually occurs. Results clearly shows that domain information is crucial for WSD, allowing our system to improve the state-of-the-art for unsupervised WSD.

The main conclusion of that work was that Semantic Domains play a dual role in linguistic description. One role is characterizing word senses (i.e. *lexical-concepts*), typically by assigning domain labels to word senses in a dictionary or lexicon. On the other hand, at a text level, Semantic Domains are clusters of texts regarding similar topics/subjects. They can be perceived as collections of domain specific texts, in which a generic corpus is organized. Examples of Semantic Domains at the text level are the subject taxonomies adopted to organize books in libraries.

The generality of these results encouraged us to extend the range of applicability of our assumptions, leading to the definition of a large number of NLP techniques relying on the common theoretical framework provided by Semantic Domains in computational linguistics (Gliozzo, 2005). For brevity, we will not describe into details all these results, limiting ourselves to enumerate the range of applicability of domain driven techniques in NLP: Word Sense Disambiguation (Gliozzo et al., 2005b), Text Categorization (Gliozzo and Strapparava, 2005b), Term Categorization (D’Avanzo et al., 2005), Ontology Learning (Gliozzo, 2006) and Multilinguality (Gliozzo and Strapparava, 2005a).

In all those tasks state-of-the-art results have been achieved by following the common methodology of acquiring Domain Models from texts by means of a common corpus based technique, inspired and motivated by the Trier’s theory and by its connection to the meaning-is-use assumption. In particular we adopted an approach based on Latent Semantic Analysis to acquire domain models from corpora describing the application domain, and we assumed the principal components so acquired be mapped to a set of semantic domains. Latent Semantic Analysis

[2] Semcor is a subportion of the Brown corpus annotated by WordNet senses.

has been performed on a term-by-document matrix capturing only co-occurrency information among terms in texts, with the aim of demonstrating our meaning-is-use assumptions. Then we exploited domain based representations to index both terms and texts, adopting a semi-supervised learning paradigm based on kernel methods. Empirical results showed that domain based representations performs better than standard bag-of-words commonly adopted for retrieval purposed, allowing a better generalization over the training data (i.e. improving the learning curve in all the supervised tasks in which they have been applyied), and allowing the definition of hybrid similarity measures to compare terms and texts, as expected from the notion of Semantic Domain.

5 Conclusion

In this paper we explicitly depicted the connections between the use of Semantic Domains in NLP and the linguistic theory motivating them. Understanding these relations provided us an useful guideline to lead our research, leading to the definition of state-of-the-art techniques for a wide range of tasks. Having in mind a clear picture of the semantic phenomena we where modeling allowed us to identify the correct applications, to predict the results of the experiments and to motivate them.

Nonetheless, several questions arises when looking at semantic domains from an epistemological point of view:

1. is the concept of semantic domain a computational theory for lexical semantics?

2. do we have enough empirical evidence to support our linguistic claims?

3. is the task based evaluation a valid epistemological framework to corroborate linguistic theory?

My personal point of view is that the task based evaluation is probably the only objective support we can provide to linguistic theory, and especially to all those issues that are more intimately related to lexical semantics. The basic motivation is that computational linguistics is also a branch of Artificial Intelligence, and then it is subjected to the behavioral Turing test. The *meaning-is-use* assumption fits perfectly this view, preventing us from applying the traditional linguistic epistemology to computational linguistics. In fact, we are interested in exploiting the language in concrete and situated linguistic games rather than representing it in an intensional way. From this point of view, the task based support we have given to our claims is a strong evidence to conclude that Semantic Domains are computational models for lexical semantics.

Anyhow, my opinion is just a minor contribution to stimulate a larger epistemological debate involving linguists, cognitive scientists, philosophers, computer scientists, engineers, among the others. I hope that my research will contribute to stimulate this debate and to find a way to escape from the "empasse" caused by the vicious distinction between empirical and theoretical methods characterizing the research in computational linguistics in the last decade.

References

E. Coseriu. 1964. Pour une sémantique diachronique structurale. *Travaux de Linguistique et de Littérature*, 2(1):139–186.

E. D'Avanzo, A. Gliozzo, and C. Strapparava. 2005. Automatic acquisition of domain information for lexical concepts. In *Proceedings of the 2nd MEANING workshop*, Trento, Italy.

F. de Saussure. 1922. *Cours de linguistique générale*. Payot, Paris.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. MIT Press.

A. Gliozzo and C. Strapparava. 2005a. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *n Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

A. Gliozzo and C. Strapparava. 2005b. Domain kernels for text categorization. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 56–63.

A. Gliozzo, C. Strapparava, and I. Dagan. 2004. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech and Language*, 18(3):275–299.

A. Gliozzo, C. Giuliano, and C. Strapparava. 2005a. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd annual meeting of the Association* *for Computational Linguistics (ACL-05)*, pages 403–410, Ann Arbor, Michigan, June.

A. Gliozzo, C. Giuliano, and C. Strapparava. 2005b. Domain kernels for words sense disambiguation. In *to appear in proc. of ACL-2005*.

A. Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, University of Trento.

A. Gliozzo. 2006. The god model. In *proceedings of EACL-06*.

S. Landes, C. Leacock, and R. I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press.

J. Lyons. 1977. *Semantics*. Cambridge University Press.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, pages 1413–1418, Athens, Greece, June.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2001. Using domain information for word sense disambiguation. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation System*, pages 111–114, Toulose, France, July.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

G. Miller. 1990. An on-line lexical database. *International Journal of Lexicography*, 13(4):235–312.

W. Porzig. 1934. Wesenhafte bedeutungsbeziehungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 58.

Procter. 1978. *Longman Dictionary of Contemporary English*.

J. Trier. 1931. *Der deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg.

J. Trier. 1934. Das sprachliche feld. eine auseinandersetzung. *Neue Fachrbücher für Wissenschaft und Jugendbildung*, 10:428–449.

S. Ullmann. 1957. *The Principles of Semantics*. Blackwell, Oxford.

L.M. Vassilyev. 1974. The theory of semantic fields: a survey. *Linguistics*, 137:79–93.

L. Weisgerber. 1939. Vom inhaltlichen aufbau des deutschen wortschatzes. *Wortfeldforschung*, pages 193–225.

L. Wittgenstein. 1965. *Philosophical Investigations*. The Macmillan Company, New York.