Building a Resource of Patterns Using Semantic Types

Octavian Popescu
FBK Research Institute
Trento, Italy
E-mail: popescu@fbk.eu

Abstract

While a word in isolation has a high potential of expressing various senses, in certain phrases this potential is restricted up to the point that one and only one sense is possible. A phrase is called sense stable if the senses of all the words compounding it do not change their sense irrespective of the context which could be added to its left or to its right. By comparing sense stable phrases we can extract corpus patterns. These patterns have slots which are filled by semantic types that capture the relevant information for disambiguation. The relationship between slots is such that a chain like disambiguation process is possible. Annotating a corpus with these kinds of patterns is beneficial for NLP, because problems such as data sparseness, noise, learning complexity are alleviated. We evaluate the inter agreement of annotators on examples coming from BNC.

Keywords: corpus patterns, semantic types, word sense disambiguation,

1. Introduction

The performances of many natural language processing (NLP) tasks, such as Word Sense Disambiguation (WSD), Textual Entailment (TE), Machine Translation (MT), can be improved by accessing a lexicon which contains information regarding the usage of words and collocating valencies of words (see among others Lin&Pantel 2001, Hanks 2005, Pustejovsky&Jezek 2008). There is a rising interest, both in theoretical and computational linguistics, in investigating what types of information must be represented, and how these types could be conveniently organized in the lexicon, in order to efficiently process natural language. In this paper we argue that it is possible to represent the relevant information for NLP tasks under the form of patterns having slots which generalize over the classes of words. We report an experiment which evaluates the agreement in finding and annotating a corpus with such patterns in two languages: English and Italian.

By patterning the behavior of verbs, it is possible to represent the interconnection between lexical knowledge and world knowledge in a computable way. A set of patterns centered on verb usage can be built on the basis of corpus evidence of word sense discriminative contexts (Popescu& Magnini 2007). Such patterns contain the exact information required to disambiguate the words occurring in phrases which are matched by one of them. In building the patterns, both syntactic and lexical semantic information is taken into account. The patterns generalize over the concrete examples found in corpus. Classes of words can be represented together by a single lexical semantic trait, usually called semantic type. A pattern indicates which semantic types are expected on which syntactic slots, in order for any of the words occurring in a phrase to be disambiguated. The extracted patterns have the property that they express a particular relationship between the senses of the words: the senses of a fraction of the words on a specific slot strictly determines the senses of the words on the rest of the slots. A consequence of this property is the fact that disambiguating only a fraction of the words, typically just one, it is possible to disambiguate the senses of all the words belonging to a pattern, through a chain-like process. This relationship is called chain clarifying relationship (CCR). Analogously, the patterns are called CCR patterns and the words found in a context matching a CCR pattern are said to observe a CCR (see section 3 for detailed examples). The semantic types are just names for classes of words which enter in a chain clarifying relationship in a verb-centered context.

In order to obtain a CCR pattern we focus on the behavior of the contexts in which the words are disambiguated. In isolation, all senses of a word are possible. By forming phrases with meaning, which naturally occur in language usage, the sense of the ambiguous words is clarified. We can actually observe not disambiguated words, but phrases which have the property that all their words are disambiguated. Such phrases have the property that the sense of their words are not affected by the larger context in which the phrase may appear. Therefore, such phrases are sense stable, in the sense that whatever context we add to the left or to the right there will be no influence on the senses of the words occurring inside the respective phrase. A sense stable phrase which contains the least possible number of words is a context which observes a CCR. In order to obtain the set of semantic types, we compare pairs of minimally different contexts which observe a CCR in which the senses of the verb are different. The difference between the words in such cases leads to the discovery of semantic types which are responsible for the senses inside a CCR pattern.

The creation of a resource of semantic types requires the examination of various phenomena mined from corpus, by means of computational methods for distributional analysis of words in context. The set of semantic types should contain information considered relevant by human readers and which corresponds to their intuition about meaning. As semantic types are directly linked to ontological categories, they should be recognizable in different languages. The inter agreement and the accurate inter-lingua alignment are, therefore, essential for reliable semantic types.

Many of the traditional NLP tasks, such as Machine Translation, Word Sense Disambiguation, Lexical Substitution and Textual Entailment, benefit from the creation of a corpus annotated with sense stable patterns which observe the CCR property. The main advantages come from the fact that such patterns: (1) are learnable, (2) provide the exact context carrying the relevant information for word senses, (3) clearly show possible entailments, (4) capture long distance phenomena, (5) help in reducing the data sparseness problem and, as we will see in Section 3.2, (6) could be modeled by a simple grammar. The information carried in such patterns is complementary to the one coming from the distributional properties of the words.

The work reported here describes an experiment of evaluating the quality and the coverage of CCR patterns together with their semantic types, considering English and Italian corpora, with the final goal of helping in creating such resources which are effective in obtaining high accuracy in NLP tasks. The focus of the work reported here is on verbs. We report the results obtained considering two sets of verbs, one in English and the other in Italian, and gathering examples from BNC (Leech&all 1994) and itWAC respectively (Baroni et al 2009).

This paper is organized as follows: we review the related works in Section 2. We present the CCR patterns methodology in Section 3, where we also discuss the complexity of annotating a corpus with CCR patterns. In Section 4 we present the guidelines for annotators and the pre-annotation work. In section 5 we describe the measures used to evaluate the inter-agreement between annotators. The paper concludes with Conclusion and Further Research section.

2. Related Works

In NLP, the context is regarded as the major source of information for the computational approaches. (Stevenson&Wilks 2000) enquire on the sources of disambiguation for words. They found that a wide range of different situations actually occur in texts, and the process of disambiguation may rely on anything from considering only part of speech to complex semantic inferences.

(Leacock&all 1993) distinguishes between topical and local context. The difference between these two types of context is made by the way the context is used: (1) for topical context, the substantive words centered around the target word, regardless of the syntactic constituency of the sentences, are the relevant information, and (2) for local context, the context is also considered from the point of view of syntactical structure, semantic distance, word order etc. While the authors present "templates" for different words, nouns, as well as the results of the "template matching process", there is no systematic inquiring into the structure of these templates and into the possibility of generalizing them.

In a series of papers Lin (Lin 1997, Lin&Pantel 2001) put forward a methodology of extracting corpus patterns based on dependency chains. They used mutual information and its derived metrics in order to compute the similarity between paths. Their patterns are restricted to subject, object positions. They maintain that a (slotX, he) is less indicative that a (slotX, sheriff). While this might be true in some cases, the measure of similarity is given by the behavior of the other components of the contexts: both *he* and *sheriff* act either exactly the same with respect to certain verb meanings, or totally differently with respect to some others. In this particular case, the common semantic type is [Human], and the relevance for word sense disambiguation of this semantic type resides in the sense discriminative patterns which differ on slot X.

The same arguments are also valid in connection with the method proposed by Li&Abe, based on MDL (Li&Abe 1998). Another limitation of these methods, which our proposal overcomes, is that they only consider subject and object positions. However, in many cases the relevant entities are adjuncts, and/or prepositions and particles. It has been shown that closed class categories, especially prepositions and particles, play an important role in disambiguation and wrong prediction are made if they are not taken into account. (see, among others, Collins and Brooks 1995, Stetina&Nagao 1997). Our results have shown that only a relatively small fraction (27%) of the extracted patterns include just the subject and/or the object.

Zhao, Meyers and Grishman (Zhao, Meyers and Grishman 2004) proposed a SVM application to slot detection, which combines two different kernels, one of them being defined on dependency trees. Their method tries to identify the possible fillers for an event, but it does not attempt to treat ambiguous cases; also, the matching score algorithm makes no distinction between the importance of the words, considering equal matching score for any word within two levels.

(Pedersen, 1998, 2005) have clustered together the examples that represent similar contexts for WSD. However, given that they adopt mainly the methodology of ordered pairs of bigrams of substantive words, their technique works only at the word level, which may lead to a data sparseness problem. Ignoring syntactic clues may increase the level of noise, as there is no control over the relevance of a bigram.

Many of the purely syntactic methods have considered the properties of the subcategorization frame of verbs. Verbs have been partitioned in semantic classes based mainly on Levin's classes of alternation. (Dorr&Jones 1996, Kipper 2000, McCarthy 2000, Lapata&Brew 2004). These semantic classes might be used in WSD via a process of alignment with hierarchies of concepts as defined in sense repository resources (Shin&Mihalcea 2005). However the problem of the consistency of alignment is still an open issue and further research must be pursued before applying these methods to WSD.

The CPA approach, (Hanks 2005-2010, Hanks&Jezek 2008), builds a large resource of corpus patterns extracted in a bottom-up manner starting from corpus. A lexicographer inspects corpus examples and through her/his intuition extracts corpus patterns which use semantic types. The CPA approach is very precise and makes available to the computational linguistics community a very informative and helpful resource. The CPA patterns are very close to the CCR patterns, but there are differences. Firstly, in CPA there is one pattern per each sense of the verb, and that pattern is paradigmatic for the usage of that particular sense of the verb. The CCR finds many patterns for the same sense of the verb and there is no attempt to categorize any of them as paradigmatic. A second difference comes from the fact that CPA patterns represent a full argument structure of the verb, while in CCR only the syntactic positions which are relevant to the sense meaning are presented. The third difference is related to the fact that there is no explicit relation between the senses of the slots of the patterns in CPA, while in CCR the very construction of the patterns is generated on the basis of this relationship. A fourth difference comes from the fact that the semantic types used in CPA denote ontological categories which express the intuition of the lexicographer. The semantic types used in CCR have no relevance on their own. However, both approaches consider patterns which generalize over the corpus examples and it is because of the fact that the two methodologies were developed independently that these differences exist. The results reported in this paper may be helpful in building a large scale evaluation of the CPA resource.

The method of extracting corpus patterns based on sense stable contexts and sense discriminative semantic types which will be presented in the next section has the advantage that is automatic, and finds patterns which could be easily constructed and managed with the available computational tools. The basic tool needed for the automatic extraction of the patterns is a dependency parser and the reported accuracy for both Italian and English is good (Manning&all 2008, Lavelli&al 2009).

3. Sense Stable Patterns using CCR

3.1 Chain Clarifying Relationship

In the first part of this section we are going to see corpus examples of context observing the CCR. The examples presented come from the BNC unless otherwise specified (the BNC label of the sentence is indicated in parenthesis).

In the following examples the different senses of the verb *see* are strictly determined by the senses of *me*, *connection*, a*nything, counselor* respectively, and vice versa, the sense of *see* determines which sense of its direct object is actuated:

ex1. (H0766) ... it 's always exciting to see a new book by him

ex2. (ABM866) we rationally see a connection between being a triangle, and having angles equal to two right angles, ...

ex.3 (HWL151) we sneaked back to see if there was anything going down.

ex.4 (FBL19) I went to see a debt counsellor

In these examples the verb *see* has different senses: *see* as *perceive with eyes* in ex1 - sense 1, *see* as *discern or deduce mentally* in ex 2 and ex3 - sense 2, *see* as *meet someone* in ex 4 - sense3. The question is two-fold: firstly, whether there is any property of the contexts of the examples ex1-ex4 which is responsible for the different senses of *see*. Secondly, in the case of a positive answer to the first question, could we express this property under the form of a pattern that is computationally manageable? In order to respond to these questions by using the examples ex1-ex4, let us first start with a set of examples which underline the relationship we are looking for.

ex5. I come to see John's photo.

ex6. I come to see John's mother.

ex7. I come to see John's problem.

Again, in the examples above, ex5-ex7, *see* has different senses: *perceive*, *meet* and *discern mentally,* respectively. We can pin point exactly what words are responsible for the differences among these examples are: *photo*, *mother* and *problem*. Therefore, there is a direct relationship between the senses of *see* and these words. This relationship is not by any means restricted only to the three words above. We can replace *photo* with p*icture, car, house* etc. in ex5, *mother* could be replaced with s*ister, boy, uncle, neighbor* etc.in ex6, and *problem* with *reason, line of thoughts* etc. in ex7. In fact, all the words which share the same lexical-semantic trait impose the same sense for *see*. These observations can be represented schematically under the shape of patterns (below, by see_* we refer to the verb *see* with sense *):

ex5P. subj=[Human] verbINF=see_1 obj=[PICTURE]

ex6P. subj=[Human] verbINF=see_2 obj=[HUMAN] ex7P. subj=[Human] verbINF=see_3 obj=[MENTAL REFLECTION]

Above, [THING], [HUMAN], [MENTAL REFLECTION] are semantic types, and subj, verbINF, obj show the syntactic function. In fact, the words *photo, mother, problem* are ambiguous themselves and in the context of examples ex5-ex7 are disambiguated by their mutual relationship with the verb *see*. There is a simple way to verify the fact that the relationship between senses of *see* and the above semantic types exists: if this relation exists indeed, then, by choosing a word characterized by two of the semantic types above, the sense of *see* is not defined. The example ex8 shows exactly this point:

ex8. I see the point.

Here, *point* could be understood as either a mark on a paper or as an individual reason, which shows that *point* is characterized by both [THING] and [MENTAL REFLECTION]. The verb *see* remains ambiguous and also the exact semantic type of *point* is not specified in example 8. If the semantic type of *point* is clarified on the basis of the information coming from a different context, then the sense of *see* is clarified as well, according to one of the patterns ex5P or ex7P. The examples ex9 and ex10 show exactly this.

ex9. I saw the point you drew.

ex10. I saw the point you raised.

In the examples ex9, ex10 the sense of *point* is disambiguated by the CCRs "[HUMAN] draw [FIGURE]" and "[HUMAN] raised [ARGUMENT]". Consequently, by the CCR property, the sense of *see* is also disambiguated in ex9 and ex10. The relationship between the syntactic and semantic types of the words in all those examples acts in a chain - like reaction between the words which are caught in phrases which are CCR. This is the the reason we call this relationship a chain clarifying relationship (CCR).
Returning to the examples ex1-ex4 we can identify the following CCRs:

ex1P verb=see_1 obj=[OBJECT]

ex2P subj=[HUMAN] verb=see_2 obj=[MENTAL REFLECTION]

ex3P subj=[HUMAN] verbINF=see_2 obj=[MENTAL REFLECTION]

ex4P subj=[HUMAN] verbINF=see_3 obj= [HUMAN]

The form of the pattern is crucial. A semantic type is relevant only on the right spot in the CCR pattern. Examples ex11, ex11P, show that the same semantic type as in ex4P plays no role in the CCR :

ex11. (H0M146) ... Lorne saw Gary as a celebrity restaurateur.

ex11P (H0M146) subj=[HUMAN] verb=see_2 obj= [HUMAN] as [SOCIAL ROLE]

The CCR patterns are not restricted to the subject and object function. The prepositional phrases are relevant for some of the CCR contexts as well. The examples below show a CCR relationship carried on by the prepositional phrase .

ex12 (ANU) ...they were treated with contempt, abandoned to poverty ...

ex12P subj=[HUMAN] verb=abandon_1 obj=[HUMAN] prepTO=[STATE]

ex13 (EFS) abandoning myself to the luxuriance of grief in libraries

ex13P subj=[HUMAN] verb=abandon_2 obj=[HUMAN] prepTO=[ATTITUDE]

The structure of a sentence may interfere with the structure of the CCRs, but cannot modify it. Even if the required slots of a CCR are not contiguous, and the verb may not be in the present tense simple, the CCR is the same, as the transformation required does not affect the CCR. However, it may be the case that the tense or aspect play a role in the CCR. For example the patterns ex5P-ex7P require the infinitive form of the verb, and this is clearly expressed by using *verbINF* , instead of *verb*, to denote the respective slot of the pattern.

The CCR based patterns do not cover entirely the whole language usage, and there are examples which involve more information for disambiguation than the one available in the syntactic structure and semantic types. However a large part of the normal language usage is disambiguated by relying on sense stable patterns based on CCRs.

There are relationships between what we considered here to be different patterns. In the following example:

ex14 I drive my daughter to her school.

ex15 I drive my daughter to her grandmother.

ex16 I drive my daughter to despair.

ex14P subj=[Human] verb=drive_1 obj=[HUMAN] prepTO=[BUILDING]

ex15P subj=[Human] verb=drive_1 obj=[HUMAN] prepTO=[HUMAN]

ex16P subj=[Human] verb=drive_2 obj=[HUMAN] prepTO=[PSYCHOLOGICAL STATE]

we have three patterns: the first two correspond to the same sense of verb *drive*, operate a vehicle - sense 1, and to compel to act in a particular way-sense 2. Focusing on the first two examples, and patterns respectively, we see that they describe the same type of event - a person being accompanied by car to a certain location. In fact, both "school" and "human" are understood as geographical points where the respective entities are located. The patterns ex14P and ex15P are variants of the pattern ex16P:

ex16P subj=[Human] verb=drive_1 obj=[HUMAN] prepTO=[GEOGRAPHICAL LOCATION]

The pattern16P is applied to ex14 and ex15, coercing *school* and *grandmother* to be understood as [GEOGRAPHICAL LOCATION], which is a semantic type that does not otherwise directly characterize either of them. The study of coercion from the perspective of sense stable patterns based on CCR is a separate subdomain. In this paper we report the results of the direct application of the CCRs[1].

In annotating a corpus with sense stable patterns based on CCR, the annotator should indicate clearly the CCR chains. The same word may belong to different CCRs, in this case the CCR which determined the sense of that particular word is marked explicitly. The study of the interaction between different CCRs is not the concern here. The annotation of words belonging to different CCRs is restricted here to the observation of semantic types and their influence on the sense of the other words in the pattern, for each individual pattern.

3.2 Complexity of CCR annotation

The importance of CCRs for WSD resides in the fact that by knowing the sense of one component, specific senses are forced for the other components. In what follows we give a formal definition of the CCR, which will help us to devise an algorithm to find CCR contexts. We start from the primitive notion of *event* (Giorgi and Pianesi, 1997). We assume that there is a set:

$$E=\{e_1, e_2, \ldots e_n\}$$

whose elements are events, and that each event can be described by a sequence of words. Let us now consider three finite sets, *W*, *S* and *G*, where:

$$W = (w_1, w_2, \ldots w_w)$$

is the set of words used to describe events in *E*,

$$S =(w_{11}, w_{12}, \ldots, w_{1m1}, w_{21}, w_{22},\ldots w_{2\,m2}, \ldots . w_{wmw})$$

is the set of words with senses, and

$$G=(g_1, g_2, \ldots g_{mg})$$

is the set of grammatical relations.

If *e* is an event described with words w_1, w_2, …w_n we assume that *e* assigns a sense $w_{i\,j}$ and a grammatical relation g_i to any of these words. Therefore we consider *e* to be the function:

$$e\text{: } P(\{w_1, w_2, \ldots w_w\}) \rightarrow (SxG)^{n}$$

$$e(w_1, w_2, \ldots w_n) = (w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \ldots w_{nin}xg_{in})$$

For a given *k* and *l*, such that $1 \leq k \leq l \leq n$, and *k* components of $e(w_1, w_2, \ldots w_n)$ we call the *chain clarifying relation* (*CCR)* of *e* the function:

$$e_{CCR}\text{: } (SxG)^{n-k} \text{ x } (WxG)^{k} \rightarrow (SxG)^{l}$$

where $e_{CCR}(w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \ldots w_{kik}xg_{ik}, w_{k+1}xg_{k+1}, w_{k+2}xg_{i2}, \ldots w_{n}xg_{in}) = (w_{1i1}, w_{2i2}, \ldots w_{lil})$

The above definition captures the intuition that in certain contexts the senses of some of the words impose a restriction on the senses of other words. When $l=n$ we have a complete sense specification, therefore the e_{CCR} function gives a sense for any of the words of *e*.

Let us consider two events *e* and *e'* such that they differ only with respect to two slots:

$e(w_1, w_2, \ldots w_n)=(w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, w_{kik}xg_{ik} \ldots w_{nin}xg_{in})$

$e'(w_{1'}, w_2, \ldots w_{n.})=(w'_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \ldots w_{kik'}xg_{ik} \ldots w_{n,in}xg_{in})$.

We infer that there is a lexical difference between w_1 and w_1' which is responsible for the sense difference between w_{kik} and $w_{kik'}$. If precisely this difference is found to be preserved for any $e(w_1,w_2,..,w_n, w_{n+1},w_{n+2},\ldots,w_m)$, then the sequence $(w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \ldots w_{kik-1}xg_{ik-1}, w_{k,ik+1}xg_{ik+1}\ldots w_{nin}xg_{in})$ is a CCR.

The apparent complexity of the grammar which generates strings which are CCRs is context dependent. Let us define a grammar , G= (N, ∑, P, S) where the set of nonterminals, N are the words of the English Language, ∑ is the set of words of the English Language with their senses, and P the production rules which associate to each phrase containing both nonterminals and terminals a string containing only terminals. The CCR property is a production rule in this grammar. Using the same notations as above, w_i and $w_{1i1}xg_{i1}$ represent the nonterminal (the word) and the nonterminal (the word with its syntactic function and a particular sense). The production rules are of the form:

$w_1, w_2, \ldots w_n \rightarrow w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, w_{kik}xg_{ik} \ldots w_{nin}xg_{in})$

$w_1, w_2, w_{kk1}xg_{k1}\ldots w_n \rightarrow w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, w_{kik}xg_{ik} \ldots w_{nin}xg_{in})$

However, considering the grammars of the CCRs rooted in verbs, the actual complexity is regular. Indeed, because of the discriminative properties related to each slot and of the possibility of using

[1] In the CPA resource some of the phenomenon related to coercion are annotated separately as "exploitation of the norm".

semantic types instead of words, the language generating only the CCR phrases is regular. The production rules are associated with each semantic type. For example, the segment of the grammar generating the ex12p and ex13P is:

abandon them to poverty → abandon them prepTO= [STATE] → abandon [HUMAN] prepTO=[STATE] → abandon_1 [HUMAN] prepTO=[STATE]

abandon myself to grief → abandon myself prepTO= [ATTITUDE] → abandon [HUMAN] prepTO= [ATTITUDE] → abandon_2 [HUMAN] prepTO= [ATTITU

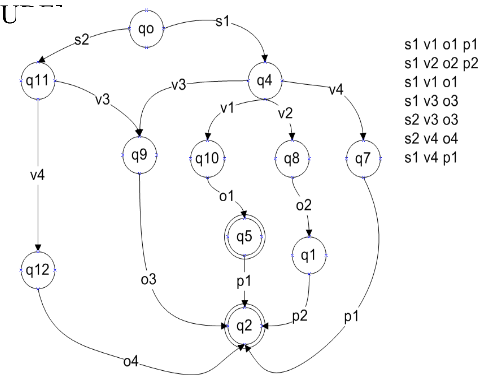

Fig 1. A regular grammar generating the CCRs

In figure 1 we plot the state transition diagram for the CCRs of a verb with the pattern set visible in the upper right corner of the figure.

The fact that what seems to be a complex characteristic of human language requiring a context dependent grammar to be modelled is actually modelled by a regular grammar has a direct impact on the annotation schema. In the following section we give the guidelines for CCR patterns annotation following the formal properties above.

4. Annotation Guidelines

The annotation guidelines of a corpus with CCRs patterns using semantic types requires an analysis of the corpus cases in order to identify the form of the semantic types. The best way to identify the semantic types is to compare the minimally different pairs, as those in example ex5-3x7. However, it is hard to find such minimal pairs in a general purpose corpus. Lacking the support of a minimally different pair, the annotator may rely on the property of sense stability for context observing CCR. Once she/he finds a sense stable context she/he may try to generalize the arguments of the verb and to determine whether there is a specific semantic type which makes sense in the respective context.

The guidelines for sense tagging on the basis of CCR are the following:

1. Identify an unambiguous phrase. To test if a phrase is unambiguous make sure that
    1.1The words have only one sense
    1.2. Part of the arguments of the verb are included
2. Check the sense stable property
    2.1 Any sub part is ambiguous - at least one word can change its meaning by adding different contexts
    2.2 Adding any other context does not change the senses of the words in the respective phrase
3. Identify the semantic types
    3.1 Replace words inside the CCR corresponding to the unambiguous phrase. If the CCR changes, discard the new word. If the CCR does not change, find the lexical feature which generalize the respective class of words (obtain the semantic type)
4. Identify the minimal elements between the CCRs that make part of the phrase.

5. Experiments

In this section we report the evaluation experiments carried on a set of verbs from Italian and English respectively. The settings and the methodology of evaluation are presented. The languages of interest are English and Italian and the sentences come from the BNC (Leech&all 1994) and itWAC, respectively (Baroni et al 2009).

The set of English and Italian verbs is presented in Table 1.

Verb	Occ	Pat	Verb	Occ	Pat
English					
begin	188	12	keep	166	19
call	208	25	move	318	36
carry	350	32	run	197	42
come	350	43	serve	112	14
develop	170	14	turn	285	28
find	150	19	use	291	21
leave	195	27	work	220	34
Italian					
arrivare	300	17	portare	300	37
accogliere	250	23	spostare	300	21
cominciare	287	26	trovare	300	32
lavorare	300	14	tacere	300	9

Table 1. English Verb Set

On the "occ" column of table 1 the number of sentences which the annotators could consult is giveb. On the "Pat" the number of patterns found is written.

Experiment

We carried the evaluation of the inter agreement in the following setting: a list of 14 English verbs, and 8 Italian verbs and 50 examples for each verb were extracted from BNC, itWac respectively. Two annotators were used for each language. The English pair are linguists:

(1) to select the context defining a pattern with semantic types,

(2) to annotate with WordNet Senses the words inside the patterns, and

(3) to specify which semantic types correspond to each slot.

The last task, (3), was performed considering three different settings involving sets of semantic types:

(3a) having the liberty to invent any semantic types according to personal intuition,

(3b) considering a predefined set of semantic types formed posteriori to (3a) by comparing the two sets of semantic types found by the two annotators and selecting semantic types from both

(3c) using a set of semantic types from the set of ontological concepts defined in SUMO (Niles, 2005), which is an ontology that is aligned with WordNet 1.6.

SUMO is an ontology which is aligned with WordNet. The concepts can be used as semantic types, thus the annotators may rely on a given set of semantic types which are also hierarchically organized. However, the SUMO categories can be used only when they are sense discriminative, in accordance with the guidelines given in section 4.

English

We measured the inter-annotator agreement on three parameters: agreement on the context considered as a pattern (ACP), word senses inside the patterns (WSP), and the equality of the semantic types (EST).

In (3a) we found ACP=98%, WSP=71%, EST=48%. In order to understand whether the EST value was due to a slight difference or to a strong disagreement, each of the annotators received the types chosen by the other one and was asked to use a true/false evaluation. It turns out that each annotator agreed with the types chosen by the other annotator in 86% and in 89% of cases, respectively. This fact suggests that the difference between the annotators was due to a lack of common denomination. In (3b) the set of types was compiled from the types found by the annotators. In order to decide which semantic types to choose, a third annotator consulted both sets of semantic types and chose the final set. The tagging of semantic types was repeated by considering this common set of types. The values for the three parameters in this second experiment are: ACP=98%, WSP=71%, ESP=93%. The value for ESP was significantly higher. In (3c) we considered a subset from the SUMO attributes as the set of predefined semantic types. The value of ESP decreased at 82% in this setting. In Table 2 we present the above figures.

ACP	WSP	EST
98%	71%	48%
98%	71%	93%
98%	71%	82%

Table 2. English Inter-agreement

We also wanted to see for the (3c) experiment how the inter-agreement is distributed for each verb individually. In Table 3 we present the percentage of the inter-agreement computed over the 50 sentences considered.

Verb	Agreement	Verb	Agreement
begin	88%	keep	90%
call	82%	move	76%
carry	76%	run	80%
come	64%	serve	90%
develop	92%	turn	88%
find	92%	use	78%
leave	82%	work	82%

Table 3. English Agreement with given Semantic Types

Italian

The same three parameters: agreement on the context considered as a pattern (ACP), word senses inside the patterns (WSP), and the equality of the semantic types (EST) were used for the Italian language as well. However, the values obtained were lower and the (3b) experiment wasn't carried yet. Notably the ACP parameter shows a much lower figure, around 85%. This could be due to the syntax of Italian, or because the two annotators weren't trained linguists. However, the agreement on the EST parameter was better, around 55%. It seems however, that the [HUMAN] semantic type was used more often than in English corpus.

ACP	WSP	EST
85%	68%	55%
85%	68%	77%

Table 2. Italian Inter-agreement

6. Conclusion and Further Research

We have presented a sense tagging methodology developed on the basis of the properties of unambiguous phrases. The words which make an unambiguous phrase are characterized by a sense chain clarifying relationship. To a chain clarifying relationship corresponds a sense discrimination pattern, which is made of lexical features. The working hypothesis is that WSD, TE, MT and other NLP systems may benefit from phrase sense relational tagging as opposed to individual and independent sense tagging.

The inter-annotator agreement suggests that it is possible to find the lexical features responsible for clarifying relationships and to maintain a consistent phrase tagging system.

We plan to extend the set of verbs for which chain clarifying relationships are determined. We also plan to experiment with a set of new features, namely, the topic words from LDOCE.

The work reported in this paper concerns only positive training examples. However, the methodology of phrase tagging allows the creation of negative examples, by using inside a CCR a wrong semantic type. We plan to develop a balanced corpus with both positive and negative examples.

In this way we can move forward towards the implementation of a full WSD system based on chain clarifying relationships methodology.

References

Baroni, M., Bernardini S., Ferraresi, A., Zanchetta, E (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 43 (3), pp: 209-226

Collins, M., Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, pp 27-38

Dorr, B., Jones, D. (1999) . Acquisition of Semantic Lexicons. In E. Viegas (ed), *Breadth and Depth of Semantic Lexicons*. Kluwer Press

Hanks, P. (2005) Immediate Context Analysis: Distinguishing Meaning by Studying usage. INn *Words in Context: A Tribute to John Sinclair on his Retirement.*

Hanks, P. (2010). How people use words to make meanings. *Natural Language Processing and Cognitive Science,* pp 3-13

Hanks, P., Jezek, E. (2010). What lexical sets tell us about lexical categories. *Corpus Linguistics and the Lexicon*, special issue of *Lexis*, *E-Journal in English Lexicology*, 4, 7-22

Lapata, M. , Brew, C. (2004). Verb Class Disambiguation Using Informative Priors. *Computational Linguistics 30:1,* pp 45-73

Lavelli, A, Hall, J., Nilsson, J., Nivre, J. (2009). MaltParser at the EVALITA 2009 Dependency Parsing Task. *EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian, 200*

Leacock, C. & al. (1993). Towards Building Contextual Represnetations of Word senses Using Statistical Models. In *Proceedings of the SIGLEX Workshop: Acquisition of Lexical KNowledge from Text*

Leech, G. Garside, R., Bryant, M. (1994). CLAWS 4: The tagging of the British National Corpus. *Proceedings of COLING 1994,* Nantes pp 622-628.

Li, D., Abe, N. (1998). Word Clustering and Disambiguation Based on Co-occurence Data . *Proceedings of* COLING-ACL, pp 749-755

Lin, D., Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering 7(4)*, pp 343-36

Lin, D. (1997). Using Syntactic Dependency as Local Context to Resolve Word sense Ambiguity. *ACL,* pp 64-7

Manning, C., D., Klein, D (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of ACL, pp 423-430*

McCarthy, D. 2000 Using semantic preferences to identify verbal participation in role switching alternation. *Proceedings of (NAACL)* , Seattle, WA

Kipper, K. , Dang, H., T., Martha Palmer. (2000). Class - Based Construction of a Verb Lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX, July 30 - August 3, 2000.

Pederson, T. (1998). Learning Probabilistic Models of Word Sense Disambiguation . Southern Methodist University (PhD Dissertation)

Pederson, T. (2005). Sense Clusters: Unsupervised Clustering and Labeling of Similar Contexts. *Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual Meeting of the Association for Computational Linguistics.*

Popescu, O., Magnini, B. (2007). Sense Discriminative Patterns for WSD, SCAR 2007, NODA, IDA 2007.

Pustejovsky, J. and Jezek, E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. *Distributional Models of the Lexicon in Linguistics and Cognitive Science.IJNL,* pp 81-214

Shin, L., Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico

Stetina, J., Nagao, M. (1997). Corpus based PP attachment ambiguity resolution with a semantic dictionary. In Zhou J, Church K (eds), *Proceedings of the 5th Workshop on very large corpora,* Beijing and Hongkong, pp 66-80.

Stevenson, K., Wilks, Y. (2001) . The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3), pp 321–349.