An Abstract Schema for Representing Semantic Roles and Modelling the Syntax-Semantics Interface

Voula Gotsoulia
University of Essex
vghotsoulia@yahoo.co.uk

Abstract

This paper presents a novel approach to semantic role annotation implementing an entailment-based view of the concept of semantic role. I propose to represent arguments of predicates with grammatically relevant primitive properties entailed by the semantics of predicates. Such meaning components generalise over a range of semantic relations which humans tend to express systematically through language. In a preliminary study, I show that we can model linguistic knowledge at a general, principled syntax-semantics interface by incorporating a layer of skeletal, entailment-based representation of word meaning in large-scale corpus annotation.

1 Introduction

Large-scale lexical semantic resources that provide *relational* information about words have recently received much focus in the field of Natural Language Processing (NLP). In particular, data-driven models for lexical semantics require the creation of broad-coverage, hand-annotated corpora with *predicate-argument* information, i.e. rich information about words expressing a semantic relation having argument slots filled by the interpretations of their grammatical complements. Corpora combining semantic and syntactic annotations constitute the backbone for the development of probabilistic models that automatically identify the semantic relationships, or *semantic roles*, conveyed by sentential constituents (Gildea and Jurafsky, 2002). That is, given an input sentence and a target *predicator* the system labels constituents with general roles like Agent, Patient, Theme, etc., or more specific roles, as in (1).

(1) [$_{Cognizer}$ I] *admired* [$_{Evaluee}$ him] [$_{Degree}$ greatly] [$_{Reason}$ for his bravery and his cheerfulness].[1]

The task of *automatic semantic role labelling* (or *shallow semantic parsing*) is a first step towards text understanding and has found use in a variety of NLP applications including information extraction (Surdeanu et al., 2003), machine translation (Boas, 2002), question answering (Narayanan and Harabagiu, 2004), summarisation (Melli et al., 2005), recognition of textual entailment relations (Burchardt and Frank, 2006), etc.

Corpora with semantic role labels additionally lend themselves to extraction of linguistic knowledge at the *syntax-semantics interface*. The range of semantic and syntactic combinatorial properties (*valences*) of each word in each of its senses is documented in terms of annotated corpus attestations. For instance, the valence pattern for the use of *admire* in (1) is shown in (2).

(2) *Cognizer*: Noun Phrase (NP), Subject
Evaluee: Noun Phrase (NP), Object
Degree: Adverbial Dependent
Reason: Prepositional Dependent

[1] This annotated example is from the FrameNet lexicon (discussed in the next section). In all examples throughout the paper, predicators are marked in italics.

This data enables the quantitative study of various linguistic phenomena and the investigation of the relationship between the distinct linguistic layers comprised by predicate-argument analysis. Furthermore, the formulation of *generalisations* over predicate-specific annotations can capture how predicates relate in terms of both semantic and syntactic features. Such syntax-semantics mappings (so-called *linking generalisations*) encode regularities concerning the associations of semantic roles with grammatical functions and are essential for a *linguistic knowledge base* for NLP applications.

This paper addresses the problem of generalising over the valences of individual predicators and proposes an abstract semantic basis for the representation of participant roles. The definition of semantic notions at an appropriate level of abstraction is the prerequisite for the formulation of a general, principled syntax-semantics interface. This is in accordance with a somewhat intuitive conception of semantic roles as classificatory notions encoding semantic similarities across different types of events or situations in the world. In effect, all conceptions of semantic roles as opposed to predicate-specific roles, such as *admirer-admired*, posit some sort of semantic classification of arguments across predicators while indicating an acknowledgment that the syntax-semantics interface (referred to with the term *linking*) is not completely arbitrary. Put differently, semantic roles constitute a level of representation suitable for capturing semantic generalisations which humans tend to express *systematically* through language.

The structure of the paper is organised as follows. Section 2 looks at conceptions of semantic roles in state-of-the-art approaches to semantic annotation indicating problems or complications related to the question of whether or how these roles can support generalisations across predicates. Section 3 calls attention to the theoretical underpinnings of the notion of semantic role and introduces an annotation schema which departs from the traditional view of semantic roles as atomic, undecomposable categories. Following the insight of Dowty's (1991) theory of Proto-Roles, I will propose analytical representations of verbal arguments based on semantically well-founded, grammatically relevant meaning components entailed by the semantics of predicates (*Proto-Role entailments*). Finally, section 4 presents a study in which lexical entailments are marked in a corpus in accordance with the proposed schema. General syntax-semantics mappings are extracted from the annotated data and are formalised in abstract classes which readily encode generalisations concerning linking to syntactic form.

2 Corpora with Semantic Roles and Related Work

Semantically annotated corpora currently available for English implement two distinct approaches to the prickly notion of semantic role. The Proposition Bank (PropBank) (Kingsbury et al., 2002) is a one million word corpus in which predicate-argument relations are hand-annotated for every occurrence of every verb in the Wall Street Journal part of the Penn Treebank (Marcus et al., 1994). Verb senses are distinguished informally on the basis of semantic as well as syntactic criteria. The semantic arguments of a verb are numbered sequentially. PropBank uses a common set of role labels (Arg0 up to Arg5) for all predicators, but these labels are defined on a per-verb basis, i.e. they have verb-specific meanings. Example PropBank annotations:

(3) a. [$_{Arg0}$ John] *broke* [$_{Arg1}$ the window] [$_{Arg2}$ with a rock].
b. [$_{Arg0}$ John] *broke* [$_{Arg1}$ the window] [$_{Arg3}$ into a million pieces].
c. [$_{Arg1}$ The window] *broke* [$_{Arg3}$ into a million pieces].

(4) [$_{Arg0}$ Blue-chip consumer stocks] *provided* [$_{Arg1}$ a lift] [$_{Arg2}$ to the industrial average].

(5) In addition, [$_{Arg0}$ the bank] has an option to *buy* [$_{Arg1}$ a 30% stake in BIP] [$_{Arg2}$ from Societe Generale] [$_{ArgM-TMP}$ after Jan.1, 1990] [$_{Arg3}$ at 1,015 francs a share].[2]

As illustrated in (3), argument labels are consistent across alternate syntactic patterns of a given predicator in a given sense. However, PropBank refrains from formalising the semantics of the role labels and does not ensure their *coherence* across verbs. This is particularly clear with higher numbered labels,

[2] *ArgM-TMP* indicates a temporal adjunct modifier.

which correspond to distinct types of participants: Arg2 marks an Instrument for *break* (3), a Benefactive for *provide* (4), and a Source for *buy* (5). Lower-numbered labels denote various roles as well, but they are less arbitrary across verbs: Arg0 corresponds to traditional Agents, Experiencers, certain types of Theme, etc. which surface as subjects of transitive verbs and a class of intransitives called unergatives; Arg1, on the other hand, is assigned to objects of transitive verbs and subjects of unaccusatives and is the equivalent of traditional Patients, Themes, etc.

While the PropBank corpus enables empirical insight into a variety of linguistic phenomena (e.g. variations in the grammatical expression of arguments) providing useful frequency information for the uses of predicators, it does not lend itself to extraction of a principled linguistic knowledge base with semantic generalisations across predicates. Inasmuch as no consistent mapping is ensured between a label and a semantic role, the argument labels result seriously overloaded across verbs. This explains why role recognition models have particularly poor performance in assigning the labels Arg2-Arg5. In fact, an attempt is currently made to map PropBank argument labels to semantically coherent roles specified by VerbNet (Kipper et al., 2000) (i.e. a broad-coverage verb lexicon based on Levin's (1993) classification of English verbs according to *shared meaning and behaviour*). Even though VerbNet specifies a small list of abstract roles (23 in total) which are intended to support generalisations, these roles are not defined as global primitives, but are meaningful only within verb classes. Because mappings of labels to semantic roles with class-specific interpretations would lead to very sparse data, argument labels are subdivided into *groupings* of VerbNet roles. The latter are created manually on the basis of analysis of argument use.[3] The subdivided (more coherent) PropBank labels perform better for semantic role labelling (Loper et al., 2007).

A different paradigm for semantic role annotation is put forth by FrameNet. The Berkeley FrameNet project (Baker et al., 1998) is creating an online lexical database containing semantic descriptions of words based on Fillmore's (1985) theory of frame semantics. The basic unit of analysis is the semantic frame, i.e. a schematic representation of a stereotypical scene or situation. Each frame is associated with a set of predicates (including verbs, nouns, and adjectives) and a set of semantic roles (called *Frame Elements*, FEs) encoding the participants and props in the designated scene. FrameNet includes manually annotated example sentences from the British National Corpus incorporating additional layers of phrase structure and grammatical function annotation. It also includes two small corpora of full-text annotation intended to facilitate statistical analysis of frame-semantic structures. Currently it contains more than 960 frames covering more than 11,600 lexical items exemplified in more than 150,000 annotated sentences. The Judgment frame evoked by *admire* in (1) is shown in Table 3.

Frame: JUDGMENT	
Definition	A Cognizer makes a judgment about an Evaluee. The judgment may be *positive* (e.g. *respect*) or *negative* (e.g. *condemn*) and this information is recorded in the semantic types Positive and Negative on the Lexical Units of this frame. There may be a specific Reason for the Cognizer's judgment, or there may be a capacity or Role in which the Evaluee is judged.
FEs	**Cognizer**: [The boss] *appreciates* you for your diligence. **Evaluee**: The boss *appreciates* [you] for your diligence. **Expressor**: She viewed him with an *appreciative* [gaze]. **Reason**: I *admire* you [for your intellect].
Predicates	accolade.n, accuse.v, admiration.n, admire.v, admiring.a, applaud.v, appreciate.v, appreciation.n, appreciative.a, approbation.n, approving.a, blame.n, blame.v, boo.v, ...

Table 1: The Judgment frame

[3]This endeavour is part of the SemLink project which aims at developing computationally explicit connections between lexical semantic resources (PropBank, VerbNet, FrameNet, WordNet). The idea is to combine the advantages of these resources and overcome their limitations by bridging the complementary lexical information they offer. In a related vein, the LIRICS (i.e. Linguistic Infrastructure for Interoperable Resources and Systems) project has recently evaluated several approaches for semantic role annotation (PropBank, VerbNet, FrameNet, among others) aiming to propose ISO ratified standards for semantic representation that will enable the exchange and reuse of (multilingual) language resources (Petukhova and Bunt, 2008).

FrameNet avoids the difficulties of attempting to pin down a small set of general roles. Instead Frame Elements are defined *locally*, i.e. in terms of frames. Frames are situated in semantic space by means of directed (asymmetric) relations. Each frame-to-frame relation associates a less dependent or more general frame (*Super_frame*) with a more dependent or less general one (*Sub_frame*). The hierarchical organisation of frames along with FE identities or analogs across frames are intended to enable the formulation of generalisations concerning the combinatorial properties (valences) of predicates. In practice, however, the frame hierarchy turns out to be somewhat complicated. Inheritance (i.e. the strongest semantic relation and the most plausible to propagate valence information across frames) is conditioned on complex sets of semantic components underlying frame definitions, ranging from FE membership and relations to other frames to relationships among FEs and Semantic Types on frames and FEs.[4] This kind of frame dependence based on fine-grained semantic or ontological distinctions is doomed to miss argument structure commonalities in predicates evoking frames that are related at a more abstract, essentially structural semantic level. Section 4 includes a concrete example of the complications in generalising valence information across FrameNet frames.

Researchers working in the FrameNet paradigm have proposed different approaches for abstracting over the properties of individual predicators and increasing the size of training data for semantic role labelling systems. Gildea and Jurafsky (2002) attempt to generalise the behaviour of semantically related predicates experimenting with a small set of abstract semantic roles mapped to FrameNet roles. Frank (2004) discusses the potential of applying various generalisation 'filters' to corpus-induced syntax-semantics mappings for abstraction of a general linguistic knowledge base. The generalisations proposed by Frank are intended to apply within frames but not across frames. Baldewein et al. (2004) have trained semantic role classifiers re-using training instances of roles that are similar to the target role. As similarity measures, they use the FrameNet hierarchy, peripheral roles of FrameNet and clusters of roles constructed automatically. Matsubayashi et al. (2009) also explore various machine learning features for generalising semantic roles in FrameNet, namely role hierarchy, human-understandable descriptors of Frame Elements, Semantic Types of filler phrases, and mappings of FrameNet roles to roles of VerbNet. The experimental result of the role classification using these generalisation features shows significant improvements in the system. This is due to the fact that role generalisations can form a remedy for the severe problem of *sparse data* which is inherent in lexical semantic corpus annotation. Data sparseness, i.e. the insufficient coverage of the range of predicate senses and constructions within sensible sizes of manually annotated data, is a bottleneck both for acquisition of linguistic knowledge for the semantic lexicon and for automated techniques for semantic role assignment.

3 An Abstract Semantic Basis for the Representation of Participant Roles

From the presentation of different annotation projects it becomes evident that semantic role annotation is a complicated task whose product is deeply influenced by its initial design philosophy and underlying criteria.[5] Among these criteria the notion of semantic role itself is central. PropBank uses general role labels that lack semantic coherence. VerbNet and FrameNet, on the other hand, specify coherent roles at a more fine-grained level (i.e. roles with class-specific or frame-specific interpretations). In this section, I consider the linguistic contours of the concept of semantic role proposing an annotation schema based upon theoretically well-founded role concepts which meet the requirements of both *generality* and *coherence*. This schema is intended at enabling the formulation of a general syntax-semantics interface suitable for modelling the relations of predicates in terms of combinatorial features.

Espousing and extending Dowty's (1991) Proto-Role hypothesis, I propose to associate arguments of predicates with properties *entailed* by the semantics of predicates.[6] Mappings of entailments to syntactic

[4] *Semantic Types* encode information that is not representable in terms of frames and FE hierarchies, e.g. basic typing of fillers of FEs referring to some (external) ontological classification, descriptions of aspects of semantic variation between lexical units such as the Positive and Negative types in the Judgment frame above, etc.

[5] This point is discussed in detail by Ellsworth et al., 2004.

[6] The term *entailment* is used in the standard logical sense according to which one formula entails another if in every possible

constituents can be many-to-one. That is, an argument can be marked with one or more properties *necessarily* entailed by the meaning of the predicator.[7] Prepositional complements are also marked with verbal entailments to which prepositions may contribute more specific content. In this paper, I will make no attempt to formalise the content added by prepositions; prepositional semantics is represented solely in terms of the common entailment basis it shares with verbal meaning.

Each Proto-Role entailment indicates a grammatically pervasive concept, i.e. a property having direct effect on the grammatical behaviour of predicates. It is defined in terms of an abstract semantic *relation* underlying the lexical meaning of the predicate. Five such relations are identified in terms of which entailment-based representations are specified: Notion, Causation, Motion, Possession, Conditioning. Note that contrary to mere ontological labels, entailment-based representations encode structural characterisations of the semantics of arguments. Consider, for instance, the sentence in (1), repeated here as (6):

(6) [$_{Cognizer}$ I] *admired* [$_{Evaluee}$ him] [$_{Reason}$ for his bravery and his cheerfulness].

A structural representation of the meaning of this construction will explicitly encode the relationships between each of the arguments of *admire*, i.e. between the NP *I* and the NP *him*, between the NP *him* and the PP *for his bravery and his cheerfulness*, and between the NP *I* and the PP *for his bravery and his cheerfulness*. By contrast, the FrameNet roles shown above do not model the fact that the semantic content of an Evaluee *requires* a Cognizer, or that a Reason *requires* both a Cognizer and an Evaluee. The view that the semantic properties underlying lexical meaning are relational in nature (i.e. they are not to be conceived entirely independently of one another) has been advocated by several researchers, among others Wechsler (1995), Pinker (1989), Jackendoff (1990), and Davis (2001), on whose work I build.

In the rest of this section, I define a set of recurring entailments which underlie the semantics of a range of verbs displaying various syntactic patterns. Note that this set can be extended on the basis of additional primitive meaning components of the sort described above, covering the semantics of broad verb classes.

(7) [$_{Conceiver}$ The other two] *pondered* [$_{Conceived}$ over this morsel] as they tramped along behind him.[8]

(8) [$_{Conceiver,Intentional}$ They] *tested* [$_{Conceived}$ the software] [$_{Conceived\_bsoa}$ for similar errors].

(9) [$_{Conceiver,Intentional}$ The government] had *reneged* [$_{Conceived}$ on promises to give them land].

(10) [$_{Conceiver}$ He] likes *stereotyping* [$_{Conceived}$ people] [$_{Conceived\_bsoa}$ by appearance].

(11) [$_{Conceiver}$ The jury] has *found out* [$_{Conceived}$ the truth] [$_{Conceived\_bsoa}$ about the suspect].

(12) [$_{Conceiver}$ The court] *categorised* [$_{Conceived,Entity}$ the issue] [$_{Conceived,Property}$ as a collateral question].

situation (in every model) in which the first is true, the second is also true. For linguistic predicates, in particular, an entailment (or lexical entailment) is an analytic implication following from the meaning of the predicate in question.

[7]The presence of 'necessarily' in this sentence is somewhat redundant, in that its meaning is incorporated by the notion of entailment. I insist, however, on emphasising it to indicate that semantic properties that are accidentally associated with the meaning of a particular *use* of a verb will not be annotated. Dowty points out that entailments of the *predicate* must be distinguished from what follows from any one sentence as a whole (e.g. entailments that may arise from NP meanings) (Dowty, 1991:572, footnote 16). For example, in the sentence *Mary slapped John*, assuming that John is a human entity, it follows from the meaning of the sentence that John will perceive something as a result of the action of slapping. But this 'entailment' is not intrinsically tied to the meaning of *slap*, because the sentences *Mary slapped the table* or *Mary slapped the corpse* are also felicitous. That is, sentience of the direct object is not an essential component of the semantics of *slap*, in the way it is for a verb like *awaken*. The sentences *Mary awakened the table* and *Mary awakened the corpse* are clearly anomalous. True entailments of predicators (which are the ones that will be annotated) must be detectable in *every possible environment* in which the predicator is used.

[8]The examples used to illustrate the proposed schema are from the British National Corpus. Some of them are slightly modified for reasons of conciseness.

(13) [$_{Conceiver}$ Opposition members] *accuse* [$_{Conceived,Entity}$ the council] [$_{Conceived,Property}$ of acting purely ideologically].

The predicates in (7)-(13) are represented in terms of a Notion relation. That is, they involve a Conceiver who is entailed to have a notion or perception of a Conceived participant (while the reverse entailment does not necessarily go through).[9] In situation types in which a Conceiver is entailed to have a notion of more than one participant, Conceived arguments are distinguished on the basis of their *salience* in the overall semantics of the predicate. For instance, *test* (8) intuitively lexicalises a dyadic relation between a Conceiver (tester) and a Conceived (tested) entity. A sought entity denoted by a *for*-PP is represented as part of a secondary Notion relation situated at the background of the primary (testing) relation. Conceived entities that are peripheral to the essential relation lexicalised by the predicate are associated with a more specific property termed *Conceived_background_state_of_affairs* (*Conceived_bsoa*). These arguments receive less *focus* in the meaning of the predicate, in a sense that they are not absolutely necessary to understand the predicate's meaning. The representation of *test* (8), *stereotype* (10), and *find out* (11) in terms of two Notion relations, one of which is treated as more salient, reifies the concept of *relative significance* of Proto-Role properties in the verbal semantics. This concept is related to the weighting of entailments in the overall semantics of a verb, which plays a critical role in determining the syntactic patterns in which the verb appears (i.e. the grammatical realisations of its arguments).[10]

The verbs in (8) and (9) involve an additional entailment of Intentionality. This is used to mark entities characterised by conscious choice, decision, or control over the course of inherently intentional actions. Intentional participants necessarily have a notion/perception of some event participant(s). The annotations in (12) and (13) include the *Entity* and *Property* tags which are intended to distinguish Conceived arguments in terms of a predicative relation assigned in the Conceiver's mental model. The *Property* label corresponds to a representation of the form *P(x)* denoting a property *P* which is predicated of some object *x*.

The entailments of Notion are not applicable in the semantics of the predicates in (14)-(15) below. These verbs refer to situations with affected participants and are described in terms of an abstract relation of Causation. In the denoted events, a Causer is entailed to affect some entity (the Causee) either physically or mentally. Causally affected participants sometimes undergo radical changes in their (physical or mental) state, which are identified in terms of a readily observable transition from a source to a final (result) state, as shown in (15).

(14) [$_{Causer}$ Diet] *influences* [$_{Causee}$ disease].

(15) [$_{Causer}$ The sun] has *changed* [$_{Causee,Change\_of\_state}$ her hair color] [$_{Source\_state}$ from red] [$_{End\_state}$ to blue].

Verbs as in (16)-(17) are represented in terms of a Motion relation involving a Moving entity (i.e. an object entailed to change location) and Stationary reference frame. Locations at the start, end, or intermediate points of the stationary frame are tagged with the labels *Path_source*, *Path_goal*, and *Path*, respectively.

(16) [$_{Moving}$ The car] *passed* [$_{Stationary}$ the railway station].

(17) [$_{Moving}$ The river] *flowed* silently [$_{Path}$ through the forest].

Finally, verbs such as *own*, *possess*, *acquire*, *lack*, etc. are treated in terms of a Possession relation involving a Possessor and an entity entailed to be Possessed (18).

[9] The Notion relation, as defined by Wechsler (1995), essentially reconstructs the entailment of *sentience*, which was proposed by Dowty (1991).

[10] Arguments identified as *conceived_bsoa*s have many of the syntactic properties of so-called semantic *adjuncts*. However, I refrain from invoking an argument versus adjunct division, in that it is known to involve serious theoretical pitfalls. Instead I classify conceived participants on the basis of the concept of importance of entailments, which lies exactly at the syntax-semantics interface. This concept is defined in terms of the *lexicalised* event rather than the real-world event that traditional analyses of adjuncthood appeal to.

(18) [$_{Possessor}$ This house] *lacks* [$_{Possessed}$ a guest room].

Verbs of caused Motion (19) or caused Possession (20) are represented in terms of both Causation and Motion/Possession, i.e. as meaning 'cause to move' (set to motion) or 'cause to possess'. This analysis posits a main (causal) event and a caused sub-event. The entailments associated with the latter are marked in square brackets.

(19) [$_{Causer}$ Lucie] *threw* [$_{Causee,[Moving]}$ him] [$_{[Path\_source]}$ from the parapet of a bridge] [$_{[Path\_goal]}$ into deep water].

(20) [$_{Causer}$ He] *handed* [$_{[Possessed]}$ the letter] [$_{[Possessor]}$ to Weir], who nodded.

Proto-Role entailments are defined in terms of inherently asymmetric semantic relations involving fixed role positions. Each of these relations (with the exception of Motion) can be thought of as instance of a more general relation entailing that properties of an entity β are dependent on an entity α. For example, a conceived entity in a Notion relation depends on the existence of a conceiver (it is taken to be within the scope of the conceiver's beliefs). An affected or possessed object in a causation or possession relation depends on the existence of some causer or possessor, respectively. I refer to this relation as Conditioning relation and associate it with appropriate Proto-Role properties capturing the semantics of a broad range of verbs for which none of the entailments specified so far seems to hold. These verbs conform to the basic transitivity pattern that motivated Dowty's Proto-Role hypothesis. Below are some characteristic examples:

(21) [$_{Condition}$ This game] *demands* [$_{Conditioned}$ great skill].

(22) [$_{Condition}$ Code 1425] *bans* [$_{Conditioned}$ large trucks in tunnels].

(23) [$_{Condition}$ The adjective 'beautiful'] *denotes* [$_{Conditioned}$ a quality which can be found in many different objects].

(24) [$_{Condition}$ Diversity] *characterises* [$_{Conditioned}$ the sociolinguistics domain].

A Conditioning relation encodes the asymmetries in such predicators in terms of the underlying entailment that the properties of a participant α impose a condition on properties of a participant β. In each of the sentences above we can conclude something about the object participant (e.g. that it is necessary, illegal, or linguistically expressed) on the basis of the subject referent (i.e. the characteristics of the game, the regulations specified by the code, the usage of the adjective 'beautiful'). By contrast, no property of the subject referent is necessarily conditioned on the object: the semantics of *ban*, for example, does not allow us to characterise code 1425 as fair/unfair, severe/lax, complete/incomplete, new/old, etc. on the basis of the object NP 'large trucks in tunnels'; similarly, we cannot infer the precise meaning of the word 'beautiful' or whether it is a verb or a noun or an adjective on the basis of the content of the NP 'a quality which can be found in many different objects'. A more precise definition of the Conditioning relation could state that the intrinsic (i.e. invariable) properties of a participant α determine or condition some non-intrinsic (i.e. variable or event-dependent) property of a participant β while the converse entailment does not go through. In (24), for example, the sociolinguistics domain is associated with a property of being diverse whereas the intrinsic properties of the domain have no significance for the definition of 'diversity' or what this notion may characterise.

4 Formulation of a General Syntax-Semantics Interface

A preliminary study has been carried out mapping state-of-the-art semantic role annotations to lexical entailment representations. In particular, a portion of the FrameNet corpora has been annotated with Proto-Role properties by a single annotator. The study focuses on a set of English verbs selected from 250 random FrameNet frames. For each verb in these frames, collections of example annotated sentences as well as sentences from the FrameNet full-text annotation corpora (where available) were extracted. More than 900 lexical units were considered in $\sim$20K sentences. Proto-Role entailments were annotated

on top of FrameNet's syntactic annotations in accordance with the schema sketched out above. The annotations were produced semi-automatically following a three-stage procedure: (i) mapping Frame Elements (FEs) to entailments at a frame level (ii) automatically adding this information to the data in a new annotation layer, (iii) manually correcting the novel annotations by examining the argument structures of individual predicators for finer semantic distinctions.

From the newly annotated data mappings of entailments to grammatical categories were acquired. The syntactic realisations of Proto-Role properties were found to readily generalise over combinatorial features of verbs pertaining to various FrameNet frames. Valence information can be formally rendered in entailment-based classes called *Lexicalisation Types* (*L-Types*) abstracting away from the semantics of predicators. L-Types are defined on the basis of grammatically relevant meaning components and encode linking generalisations cutting across FrameNet frames.

For instance, predicates such as *believe* and *desire* (evoking the frames Religious_Belief and Desiring, respectively) involve arguments that are equivalent in terms of entailments, as illustrated in (25)-(26) below. Hence they are categorised in the Notion L-Type shown in Table 2. Table 2 includes the correspondences between combinations of entailments and FrameNet Frame Elements.

Notion L-Type	Religious_belief	Desiring
Conceiver	Believer	Experiencer
Conceived, (Entity)	Element	Focal_participant
Conceived_bsoa, Property	Role	Role_of_focal_participant

Table 2: Mappings between Notion L-Type and FrameNet frames

(25) If [$_{Conceiver}$ he] *believes* [$_{Conceived,Entity}$ in Jesus] [$_{Conceived\_bsoa,Property}$ as his Saviour], he can be baptised.

(26) [$_{Conceiver}$ He] *wanted* [$_{Conceived,Entity}$ Smith] [$_{Conceived\_bsoa,Property}$ as the new producer].

In a similar fashion, *operate*, *research*, and *ratify* can be grouped together in a L-Type based on the underlying property of Intentionality. Examples (27)-(28) show that these verbs share common valence patterns despite the differences in the definition of the frames they evoke (Using, Research and Ratification): Role and Purpose are core Frame Elements in the Using frame, while Purpose is peripheral in Research and Ratification. Research and Ratification have no Role FE (but this kind of argument is clearly present in the constructions exemplified in (28b-c).

Intentionality L-Type	Using	Research	Ratification
Conceiver, Intentional	Agent	Researcher	Ratifier
Conceived, (Entity)	Instrument	Question	Proposal
Conceived_bsoa, Property	Role		
Conceived_bsoa, Intention	Purpose	Purpose	Purpose

Table 3: Mappings between Intentionality L-Type and FrameNet frames

(27) a. [$_{Conceiver,Intentional}$ We] *operate* [$_{Conceived}$ a menu] [$_{Conceived\_bsoa,Intention}$ to get the best out of rations].

b. [$_{Conceiver,Intentional}$ We] *research* [$_{Conceived}$ this fungus] [$_{Conceived\_bsoa,Intention}$ to fight aliments in tobacco and tomato fields].

c. [$_{Conceiver,Intentional}$ They] had to *ratify* [$_{Conceived}$ the amendments] [$_{Conceived\_bsoa,Intention}$ to be readmitted to the Union].

(28) a. There has been a long debate as to whether [$_{Conceived,Entity}$ the Severn Mill] *was operated* [$_{Conceived\_bsoa,Property}$ as a tide mill].

b. [$_{Conceived,Entity}$ Thin films] *are being researched* [$_{Conceived\_bsoa,Property}$ as a potential medium for integrated optical circuits].

c. [$_{Conceived,Entity}$ Such agreements] may *be ratified* [$_{Conceived\_bsoa,Property}$ as being in the public interest].

In the same Intentionality L-Type we also categorise verbs such as *carry out* and *visit* evoking the frames Intentionally_act and Visiting. It is important to note that despite the argument structure similarities of these predicators, it is not possible to establish an identity link between the Act FE of the Intentionally_act frame and the Entity FE of Visiting in terms of the frame hierarchy, because the FEs are associated with different Semantic Types in the corresponding frame definitions, i.e. Act is of type State_of_affairs whereas Entity is of type Physical_object. The examples (29)-(30) illustrate the common use of these verbs in the transitive construction. The (a) sentences show the FE annotation while the (b) sentences show the annotated entailments.

(29) a. [$_{Agent}$ They] had *carried out* [$_{Act}$ 113 uranium conversion experiments].
b. [$_{Conceiver,Intentional}$ They] had *carried out* [$_{Conceived}$ 113 uranium conversion experiments].

(30) a. [$_{Agent}$ You] have to *visit* [$_{Entity}$ your parents] every once in a while.
b. [$_{Conceiver,Intentional}$ You] have to *visit* [$_{Conceived}$ your parents] every once in a while.

Predicates grouped together in L-Types have some but not necessarily all their grammatical properties in common. This is in accordance with the fact that L-Types are essentially semantically-driven modelling recurring, abstract features in the semantics of predicators while disregarding ephemeral properties as well as lexical idiosyncrasies.[11] In addition to the set of entailments discussed in the previous section, L-Types may also incorporate more fine-grained properties that are clearly relevant to linking. For instance, verbs lexicalising a Desiring situation were found with prepositional complements introduced by *for*, *after*, *to*, *towards*, *of*, or *over* (e.g. *long for*, *hanker after*, *aspire to*, *pine over*, etc.), but not *on*, *upon*, *at*, or *about* (like other Notion verbs, such as *ponder*, *muse*, *think*, etc.). Inasmuch as a Desiring relation is identified as a recurring concept systematically associated with a particular grammatical relation (e.g. a *for*-PP), it can be represented in a separate L-Type inheriting from the Notion L-Type presented previously.[12] An initial classification like the one exemplified above captures general conditions which determine possible associations between the semantics of predicators and grammatical relations realising their arguments (e.g. the fact that a conceived entity can only surface in subject position in a passive sentence). It can be extended and refined on the basis of more specific semantic relations. Moreover, L-Types can be organised in hierarchical structures. They can form the upper portion of a principled hierarchy of classes encoding successively broader levels of generalisations concerning argument linking.

This study indicated that a small number of Lexicalisation Types abstracts over a wide range of FrameNet frames.[13] More precisely, in the annotated dataset 48 L-Types were identified based on various combinations of entailments: 9 Notion Types, 7 Intentionality Types, 10 Causation Types, 7 Communication (Caused_Notion) Types, 7 Motion (including Caused_Motion) Types, 7 Possession (including Caused_Possession) Types, and 1 Conditioning Type. These Types readily abstract over associations of semantic properties and grammatical functions attested in over 200 FrameNet frames.[14] In the FrameNet paradigm, L-Types can be modelled as non-lexicalised frames specifying syntactic mapping constraints.

[11] L-Types crucially differ from verb classes in VerbNet, which are based on a rigorous commitment to syntax. This commitment yields fine-grained distinctions that very often split semantically coherent classes. In fact, L-Types abstract over VerbNet classes encoding broader levels of linking generalisations.

[12] *For*-PPs are indeed associated with a desiderative sense with a wide range of verbs in various argument positions: 'He desperately hunted *for a new job*'. 'They searched the ground *for traces*'. 'John ran *for cover* when it started to rain'.

[13] Note that inasmuch as L-Types abstract over both VerbNet classes and FrameNet frames, they can also be useful for combining the two resources.

[14] About 30 frames contained predicates for which none of our entailments seemed to hold. Most of these verbs (e.g. *resemble*, *adjoin*, *concern*, *fit*, *suit*, etc.) involve what Dowty (1991) called *perspective-dependent* semantic roles traditionally described with labels such as Figure and Ground. The lexicalisation patterns of these verbs have been shown to depend on *pragmatic* or *discourse* factors rather than intrinsic semantic properties. Such predicates display great variability in their argument realisation options and are outside the scope of this study.

Mappings between FrameNet frames and L-Types can be stated by means of a separate relation in addition to the frame relations currently specified by FrameNet. A relation generalising the combinatorial properties of lexical items across frames would simplify the picture of the frame hierarchy, in that it would essentially decouple purely lexical semantic information (encoded by existing frame-to-frame relations) from information pertaining exactly to the interface of syntax and semantics. In future work, our intention is to test whether the proposed semantic role schema and the attested L-Types can be useful for dealing with the sparse data problem and increasing the performance of semantic role labelling systems.

References

[1] Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Proceedings of the COLING-ACL. Montreal, Canada.

[2] Baldewein, Ulrike, Katrin Erk, Sebastian Padó, and Detlef Prescher. 2004. Semantic Role Labeling with Similarity-Based Generalisation Using EM-Based Clustering. In Proceedings of Senseval-3, pp. 64-68. Barcelona, Spain.

[3] Boas, Hans C. 2002. Bilingual FrameNet Dictionaries for Machine Translation. In Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain. Vol. IV: pp. 1364-1371.

[4] Burchardt, Aljoscha and Anette Frank. 2006. Approximating Textual Entailment with LFG and FrameNet Frames. In Proceedings of the second PASCAL Recognizing Textual Entailment Workshop. Venice, Italy, pp. 92-97.

[5] Davis, Anthony. 2001. Linking by types in the hierarchical lexicon. CSLI Publications.

[6] Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. Language 67.3, pp. 547-619.

[7] Ellsworth Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. PropBank, SALSA, and FrameNet: How Design Determines Product. In Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon.

[8] Fillmore, Charles J. 1985. Frames and the semantics of understanding. Quaderni di Semantica 6.2, pp. 222-254.

[9] Frank, Anette. 2004. Generalizations over corpus-induced frame assignment rules. In Charles Fillmore, Manfred Pinkal, Collin Baker and Katrin Erk (eds.): Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon, Portugal, pp. 31-38.

[10] Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics 28 (3), pp. 245-288.

[11] Jackendoff, Ray. 1990. Semantic Structures. Cambridge, MA, MIT Press.

[12] Kingsbury, Paul and Martha Palmer. 2002. From Treebank to PropBank. In Proceedings of the LREC, Las Palmas, Canary Islands, Spain.

[13] Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), Austin,TX, July-August.

[14] Levin, Beth. 1993. English verb classes and alternations: A preliminary investigation. University of Chicago Press.

[15] Loper, Edward, Szu-ting Yi and Martha Palmer. 2007. Combining Lexical Resources: Mapping Between PropBank and VerbNet. Proceedings of the 7th International Workshop on Computational Semantics. Tilburg, the Netherlands.

[16] Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In Proceedings AR-PAHLT Workshop.

[17] Matsubayashi, Yuichiroh, Naoaki Okazaki, and Jun'ichi Tsujii. 2009. A Comparative Study on Generalization of Semantic Roles in FrameNet. In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp.19-27. Suntec, Singapore.

[18] Melli, Gabor, Yang Wang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar, and Fred Popowich. 2005. Description of SQUASH, the SFU question answering summary handler for the DUC-2005 Summarization Task. In Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC). Vancouver, Canada, available at http://duc.nist.gov/pubs/2005papers/simonfraseru.sarkar.pdf.

[19] Narayanan, Srini and Sanda Harabagiu. 2004. Question Answering Based on Semantic Structures. In Proceedings of the 20th International Conference on Computational Linguistics (COLING), pp. 693 - 701. Geneva, Switzerland.

[20] Petukhova, Volha and Harry Bunt. 2008. 'LIRICS semantic role annotation: design and evaluation of a set of data categories.' In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May 28-30.

[21] Pinker, Steven. 1989. Learnability and Cognition. Cambridge, MA, MIT Press.

[22] Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate Arguments Structures for Information Extraction. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 815. Sapporo, Japan.

[23] Wechsler, Stephen. 1995. The semantic basis of argument structure. Stanford, CA. CSLI Publications.