Semantic Role Labeling: An Introduction to the Special Issue

Lluís Màrquez*
Universitat Politècnica de Catalunya

Xavier Carreras**
Massachusetts Institute of Technology

Kenneth C. Litkowski†
CL Research

Suzanne Stevenson‡
University of Toronto

Semantic role labeling, the computational identification and labeling of arguments in text, has become a leading task in computational linguistics today. Although the issues for this task have been studied for decades, the availability of large resources and the development of statistical machine learning methods have heightened the amount of effort in this field. This special issue presents selected and representative work in the field. This overview describes linguistic background of the problem, the movement from linguistic theories to computational practice, the major resources that are being used, an overview of steps taken in computational systems, and a description of the key issues and results in semantic role labeling (as revealed in several international evaluations). We assess weaknesses in semantic role labeling and identify important challenges facing the field. Overall, the opportunities and the potential for useful further research in semantic role labeling are considerable.

1. Introduction

The sentence-level semantic analysis of text is concerned with the characterization of events, such as determining "who" did "what" to "whom," "where," "when," and "how." The predicate of a clause (typically a verb) establishes "what" took place, and other sentence constituents express the participants in the event (such as "who" and "where"), as well as further event properties (such as "when" and "how"). The primary task of **semantic role labeling** (SRL) is to indicate exactly what semantic relations hold among a predicate and its associated participants and properties, with these relations

* Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Jordi Girona Salgado 1–3, 08034 Barcelona, Spain. E-mail: lluism@lsi.upc.edu.
** Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, 32 Vassar St., Cambridge, MA 02139, USA. E-mail: carreras@csail.mit.edu.
† CL Research, 9208 Gue Road, Damascus, MD 20872 USA. E-mail: ken@clres.com.
‡ Department of Computer Science, 6 King's College Road, Toronto, ON M5S 3G4, Canada. E-mail: suzanne@cs.toronto.edu.

drawn from a pre-specified list of possible **semantic roles** for that predicate (or class of predicates). In order to accomplish this, the role-bearing constituents in a clause must be identified and their correct semantic role labels assigned, as in:

[The girl on the swing]$_{Agent}$ [whispered]$_{Pred}$ to [the boy beside her]$_{Recipient}$

Typical roles used in SRL are labels such as Agent, Patient, and Location for the entities participating in an event, and Temporal and Manner for the characterization of other aspects of the event or participant relations. This type of role labeling thus yields a first-level semantic representation of the text that indicates the basic event properties and relations among relevant entities that are expressed in the sentence.

Research has proceeded for decades on manually created lexicons, grammars, and other semantic resources (Hirst 1987; Pustejovsky 1995; Copestake and Flickinger 2000) in support of deep semantic analysis of language input, but such approaches have been labor-intensive and often restricted to narrow domains. The 1990s saw a growth in the development of statistical machine learning methods across the field of computational linguistics, enabling systems to learn complex linguistic knowledge rather than requiring manual encoding. These methods were shown to be effective in acquiring knowledge necessary for semantic interpretation, such as the properties of predicates and the relations to their arguments—for example, learning subcategorization frames (Briscoe and Carroll 1997) or classifying verbs according to argument structure properties (Merlo and Stevenson 2001; Schulte im Walde 2006). Recently, medium-to-large corpora have been manually annotated with semantic roles in FrameNet (Fillmore, Ruppenhofer, and Baker 2004), PropBank (Palmer, Gildea, and Kingsbury 2005), and NomBank (Meyers et al. 2004), enabling the development of statistical approaches specifically for SRL.

With the advent of supporting resources, SRL has become a well-defined task with a substantial body of work and comparative evaluation (see, among others, Gildea and Jurafsky [2002], Surdeanu et al. [2003], Xue and Palmer [2004], Pradhan et al. [2005a], the CoNLL Shared Task in 2004 and 2005, and Senseval-3 and SemEval-2007). The identification of event frames may potentially benefit many natural language processing (NLP) applications, such as information extraction (Surdeanu et al. 2003), question answering (Narayanan and Harabagiu 2004), summarization (Melli et al. 2005), and machine translation (Boas 2002). Related work on classifying the semantic relations in noun phrases has also been encouraging for NLP tasks (Moldovan et al. 2004; Rosario and Hearst 2004).

Although the use of SRL systems in real-world applications has thus far been limited, the outlook is promising for extending this type of analysis to many applications requiring some level of semantic interpretation. SRL represents an excellent framework with which to perform research on computational techniques for acquiring and exploiting semantic relations among the different components of a text.

This special issue of *Computational Linguistics* presents several articles representing the state-of-the-art in SRL, and this overview is intended to provide a broader context for that work. First, we briefly discuss some of the linguistic views on semantic roles that have had the most influence on computational approaches to SRL and related NLP tasks. Next, we show how the linguistic notions have influenced the development of resources that support SRL. We then provide an overview of SRL methods and describe the state-of-the-art as well as current open problems in the field.

2. Semantic Roles in Linguistics

Since the foundational work of Fillmore (1968), considerable linguistic research has been devoted to the nature of semantic roles. Although there is substantial agreement on major semantic roles, such as Agent and Theme, there is no consensus on a definitive list of semantic roles, or even whether such a list exists. Proposed lists range from a large set of situation-specific roles, such as Suspect, Authorities, and Offense (Fillmore, Ruppenhofer, and Baker 2004), to a relatively small set of general roles, such as Agent, Theme, Location, and Goal (typically referred to as **thematic roles**, as in Jackendoff [1990]), to the set of two core roles, Proto-Agent and Proto-Theme, whose entailments determine the precise relation expressed (Dowty 1991). This uncertainty within linguistic theory carries over into computational work on SRL, where there is much variability on the roles assumed in different resources.

A major focus of work in the linguistics community is on the mapping between the predicate–argument structure that determines the roles, and the syntactic realization of the recipients of those roles (Grimshaw 1990; Levin 1993; Levin and Rappaport Hovav 2005). Semantic role lists are generally viewed as inadequate for explaining the morphosyntactic behavior of argument expression, with argument realization dependent on a deeper lexical semantic representation of the components of the event that the predicate describes. Although much of the mapping from argument structure to syntax is predictable, this mapping is not completely regular, nor entirely understood. An important question for SRL, therefore, is the extent to which performance is degraded by the irregularities noted in linguistic studies of semantic roles.

Nonetheless, sufficient regularity exists to provide the foundation for meaningful generalizations. Much research has focused on explaining the varied expression of verb arguments within syntactic positions (Levin 1993). A major conclusion of that work is that the patterns of syntactic alternation exhibit regularity that reflects an underlying semantic similarity among verbs, forming the basis for verb classes. Such classes, and the argument structure specifications for them, have proven useful in a number of NLP tasks (Habash, Dorr, and Traum 2003; Shi and Mihalcea 2005), including SRL (Swier and Stevenson 2004), and have provided the foundation for the computational verb lexicon VerbNet (Kipper, Dang, and Palmer 2000).

This approach to argument realization focuses on the relation of morphosyntactic behavior to argument semantics, and typically leads to a general conceptualization of semantic roles. In frame semantics (Fillmore 1976), on the other hand, a word activates a frame of semantic knowledge that relates linguistic semantics to encyclopedic knowledge. This effort has tended to focus on the delineation of situation-specific frames (e.g., an Arrest frame) and correspondingly more specific semantic roles (e.g., Suspect and Authorities) that codify the conceptual structure associated with lexical items (Fillmore, Ruppenhofer, and Baker 2004). With a recognition that many lexical items could activate any such frame, this approach leads to lexical classes of a somewhat different nature than those of Levin (1993). Whereas lexical items in a Levin class are syntactically homogeneous and share coarse semantic properties, items in a frame may syntactically vary somewhat but share fine-grained, real-world semantic properties.

A further difference in these perspectives is the view of the roles themselves. In defining verb classes that capture argument structure similarities, Levin (1993) does not explicitly draw on the notion of semantic role, instead basing the classes on behavior that is hypothesized to reflect the properties of those roles. Other work also eschews the notion of a simple list of roles, instead postulating underlying semantic structure that captures the relevant properties (Levin and Rappaport Hovav 1998). Interestingly,

as described in Fillmore, Ruppenhofer, and Baker (2004), frame semantics also avoids a predefined list of roles, but for different reasons. The set of semantic roles, called **frame elements**, are chosen for each frame, rather than being selected from a predefined list that may not capture the relevant distinctions in that particular situation. Clearly, to the extent that disagreement persists on semantic role lists and the nature of the roles themselves, SRL may be working on a shifting target.

These approaches also differ in the broad characterization of event participants (and their roles) as more or less essential to the predicate. In the more syntactic-oriented approaches, roles are typically divided into two categories: **arguments**, which capture a core relation, and **adjuncts**, which are less central. In frame semantics, the roles are divided into **core** frame elements (e.g., Suspect, Authorities, Offense) and **peripheral** or **extra-thematic** elements (e.g., Manner, Time, Place). These distinctions carry over into SRL, where we see that systems generally perform better on the more central arguments.

Finally, although predicates are typically expressed as verbs, and thus much work in both linguistics and SRL focuses on them, some nouns and adjectives may be used predicatively, assigning their own roles to entities (as in the adjective phrase *proud that we finished the paper*, where the subordinate clause is a Theme argument of the adjective *proud*). Frame semantics tends to include in a frame relevant non-verb lexical items, due to the emphasis on a common situation semantics. In contrast, the morphosyntactic approaches have focused on defining classes of verbs only, because they depend on common syntactic behavior that may not be apparent across syntactic categories.

Interestingly, prepositions have a somewhat dual status with regard to role labeling. In languages like English, prepositions serve an important function in signaling the relation of a participant to a verb. For example, it is widely accepted that *to* in *give the book to Mary* serves as a grammatical indicator of the Recipient role assigned by the verb, rather than as a role assigner itself. In other situations, however, a preposition can be viewed as a role-assigning predicate in its own right. Although some work in computational linguistics is tackling the issue of the appropriate characterization of prepositions and their contribution to semantic role assignment (as we see subsequently), much work remains in order to fully integrate linguistic theories of prepositional function and semantics into SRL.

3. From Linguistic Theory to Computational Resources

The linguistic approaches to semantic roles discussed previously have greatly influenced current work on SRL, leading to the creation of significant computational lexicons capturing the foundational properties of predicate–argument relations.

In the FrameNet project (Fillmore, Ruppenhofer, and Baker 2004), lexicographers define a frame to capture some semantic situation (e.g., Arrest), identify lexical items as belonging to the frame (e.g., *apprehend* and *bust*), and devise appropriate roles for the frame (e.g., Suspect, Authorities, Offense). They then select and annotate example sentences from the British National Corpus and other sources to illustrate the range of possible assignments of roles to sentence constituents for each lexical item (at present, over 141,000 sentences have been annotated).

FrameNet thus consists of both a computational lexicon and a role-annotated corpus. The existence of such a corpus enabled Gildea and Jurafsky (2002) to develop the first statistical machine learning approach to SRL, using various lexical and syntactic features such as phrase type and grammatical function calculated over the annotated constituents. Although this research spurred the current wave of SRL work that has

refined and extended Gildea and Jurafsky's approach, the FrameNet data has not been used extensively. One issue is that the corpus is not a representative sample of the language, but rather consists of sentences chosen manually to illustrate the possible role assignments for a given lexical item. Another issue is that the semantic roles are situation-specific, rather than general roles like Agent, Theme, and Location that can be used across many situations and genres.

The computational verb lexicon, VerbNet (Kipper, Dang, and Palmer 2000), instead builds on Levin's (1993) work on defining verb classes according to shared argument realization patterns. VerbNet regularizes and extends the original Levin classes; moreover, each class is explicitly associated with argument realization specifications that state the constituents that a verb can occur with and the role assigned to each. The roles are mostly drawn from a small set (around 25) of general roles widely used in linguistic theory. This lexicon has been an important resource in computational linguistics, but because of the lack of an associated role-annotated corpus, it has only been used directly in SRL in an unsupervised setting (Swier and Stevenson 2004).

Research on VerbNet inspired the development of the Proposition Bank (PropBank; Palmer, Gildea, and Kingsbury 2005), which has emerged as a primary resource for research in SRL (and used in four of the articles in this special issue). PropBank addresses some of the issues for SRL posed by the FrameNet data. First, the PropBank project has annotated the semantic roles for all verbs in the Penn Treebank corpus (the *Wall Street Journal* [WSJ] news corpus). This provides a representative sample of text with role-annotations, in contrast to FrameNet's reliance on manually selected, illustrative sentences. Importantly, PropBank's composition allows for consideration of the statistical patterns across natural text. Although there is some concern about the limited genre of its newspaper text, this aspect has the advantage of allowing SRL systems to benefit from the state-of-the-art syntactic parsers and other resources developed with the WSJ TreeBank data. Moreover, current work is extending the PropBank annotation to balanced corpora such as the Brown corpus.

The lexical information associated with verbs in PropBank also differs significantly from the situation-specific roles of FrameNet. At the same time, the PropBank designers recognize the difficulty of providing a small, predefined list of semantic roles that is sufficient for all verbs and predicate–argument relations, as in VerbNet. PropBank instead takes a "theory-neutral" approach to the designation of core semantic roles. Each verb has a **frameset** listing its allowed role labelings in which the arguments are designated by number (starting from 0). Each numbered argument is provided with an English-language description specific to that verb. Participants typically considered as adjuncts are given named argument roles, because there is more general agreement on such modifiers as Temporal or Manner applying consistently across verbs. Different senses for a polysemous verb have different framesets; however, syntactic alternations which preserve meaning (as identified in Levin [1993]) are considered to be a single frameset. While the designations of Arg0 and Arg1 are intended to indicate the general roles of Agent and Theme/Patient across verbs, other argument numbers do not consistently correspond to general (non-verb-specific) semantic roles.

Given the variability in the sets of roles used across the computational resources, an important issue is the extent to which different role sets affect the SRL task, as well as subsequent use of the output in other NLP applications. Gildea and Jurafsky (2002) initiated this type of investigation by exploring whether their results were dependent on the set of semantic roles they used. To this end, they mapped the FrameNet frame elements into a set of *abstract thematic roles* (i.e., more general roles such as Agent, Theme, Location), and concluded that their system could use these thematic roles

without degradation. Similar questions must be investigated in the context of PropBank, where the framesets for the verbs may have significant domain-specific meanings and arguments due to the dependence of the project on WSJ data. Given the uncertainty in the linguistic status of semantic role lists, and the lack of evidence about which types of roles would be most useful in various NLP tasks, an important ongoing focus of attention is the value of mapping between the role sets of the different resources (Swier and Stevenson 2005; Loper, Yi, and Palmer 2007; Yi, Loper, and Palmer 2007).

We noted previously the somewhat special part that prepositions play in marking semantic relations, in some sense mediating the role assignment of a verb to an argument. The resources noted earlier differ in their treatment of prepositions. In VerbNet, for example, prepositions are listed explicitly as part of the syntactic context in which a role is assigned (e.g., *Agent V Prep(for) Recipient*), but it is the NP object of the preposition that receives the semantic role. In FrameNet and PropBank, on the other hand, the full prepositional phrase is considered as the frame element (the constituent receiving the role). Clearly, further work needs to proceed on how to best capture the interaction between verbs and prepositions in SRL. This is especially complex given the high polysemy of prepositions, and work has proceeded on relating preposition disambiguation to role assignment (e.g., O'Hara and Wiebe 2003). For such approaches to make meaningful progress, resources are needed that elaborate the senses of prepositions and relate those senses to semantic roles. In The Preposition Project (TPP; Litkowski and Hargraves 2005), a comprehensive, hierarchical characterization of the semantic roles for all preposition senses in English is being developed. TPP has sense-tagged more than 25,000 preposition instances in FrameNet sentences, allowing for comprehensive investigation of the linking between preposition sense and semantic role assignment.

4. Approaches to Automatic SRL

The work on SRL has included a broad spectrum of probabilistic and machine-learning approaches to the task. We focus here on supervised systems, because most SRL research takes an approach requiring training on role-annotated data. We briefly survey the main approaches to automatic SRL, and the types of learning features used.

4.1 SRL Step by Step

Given a sentence and a designated verb, the SRL task consists of identifying the boundaries of the arguments of the verb predicate (argument identification) and labeling them with semantic roles (argument classification). The most common architecture for automatic SRL consists of the following steps to achieve these subtasks.

The first step in SRL typically consists of **filtering** (or pruning) the set of argument candidates for a given predicate. Because arguments may be a continuous or discontinuous sequence of words, any subsequence of words in the sentence is an argument candidate. Exhaustive exploration of this space of candidates is not feasible, because it is both very large and imbalanced (i.e., the vast majority of candidates are not actual arguments of the verb). The simple heuristic rules of Xue and Palmer (2004) are commonly used to perform filtering because they greatly reduce the set of candidate arguments, while maintaining a very high recall.

The second step consists of a **local scoring** of argument candidates by means of a function that outputs probabilities (or confidence scores) for each of the possible

role labels, plus an extra "no-argument" label indicating that the candidate should not be considered an argument in the solution. In this step, candidates are usually treated independently of each other. A crucial aspect in local scoring (see Section 4.2) is the representation of candidates with features, rather than the particular choice of classification algorithm.

Argument **identification** and **classification** may be treated jointly or separately in the local scoring step. In the latter case, a pipeline of two subprocesses is typically applied, first scoring between "argument" and "no-argument" labels, and then scoring the particular argument labels. Because argument identification is closely related to syntax and argument classification is more a semantic issue, useful features for the two subtasks may be very different—that is, a good feature for addressing recognition may hurt classification and vice versa (Pradhan et al. 2005a).

The third step in SRL is to apply a **joint scoring** (or global scoring) in order to combine the predictions of local scorers to produce a good structure of labeled arguments for the predicate. In this step, dependencies among several arguments of the same predicate can be exploited. For instance, Punyakanok, Roth, and Yih (this issue) ensure that a labeling satisfies a set of structural and SRL-dependent constraints (arguments do not overlap, core arguments do not repeat, etc.). Also in this issue, Toutanova, Haghighi, and Manning apply re-ranking to select the best among a set of candidate complete solutions produced by a base SRL system. Finally, probabilistic models have also been applied to produce the structured output, for example, generative models (Thompson, Levy, and Manning 2003), sequence tagging with classifiers (Màrquez et al. 2005; Pradhan et al. 2005b), and Conditional Random Fields on tree structures (Cohn and Blunsom 2005). These approaches at a global level may demand considerable extra computation, but current optimization techniques help solve them quite efficiently.

Some variations in the three-step architecture are found. Systems may bypass one of the steps, by doing only local scoring, or skipping directly to joint scoring. A fourth step may consist of fixing common errors or enforcing coherence in the final solution. This postprocess usually consists of a set of hand-developed heuristic rules that are dependent on a particular architecture and corpus of application.

An important consideration within this general SRL architecture is the *combination* of systems and input annotations. Most SRL systems include some kind of combination to increase robustness, gain coverage, and reduce effects of parse errors. One may combine: (1) the output of several independent SRL basic systems (Surdeanu et al. 2007; Pradhan et al. 2005b), or (2) several outputs from the same SRL system obtained by changing input annotations or other internal parameters (Koomen et al. 2005; Toutanova, Haghighi, and Manning 2005). The combination can be as simple as selecting the best among the set of complete candidate solutions, but usually consists of combining fragments of alternative solutions to construct the final output. Finally, the combination component may involve machine learning or not. The gain in performance from the combination step is consistently between two and three F_1 points. However, a combination approach increases system complexity and penalizes efficiency.

Several exceptions to this described architecture for SRL can be found in the literature. One approach entails joint labeling of all predicates of the sentence, instead of proceeding one by one. This opens the possibility of exploiting dependencies among the different verbs in the sentence. However, the complexity may grow significantly, and results so far are inconclusive (Carreras, Màrquez, and Chrupała 2004; Surdeanu et al. 2007). Other promising approaches draw on dependency parsing rather than traditional phrase structure parsing (Johansson and Nugues 2007), or combine parsing and SRL into a single step of semantic parsing (Musillo and Merlo 2006).

4.2 Feature Engineering

As previously noted, devising the features with which to encode candidate arguments is crucial for obtaining good results in the SRL task. Given a verb and a candidate argument (a syntactic phrase) to be classified in the local scoring step, three types of features are typically used: (1) features that characterize the candidate argument and its context; (2) features that characterize the verb predicate and its context; and (3) features that capture the relation (either syntactic or semantic) between the candidate and the predicate.

Gildea and Jurafsky (2002) presented a compact set of features across these three types, which has served as the core of most of the subsequent SRL work: (1) the phrase type, headword, and governing category of the constituent; (2) the lemma, voice, and subcategorization pattern of the verb; and (3) the left/right position of the constituent with respect to the verb, and the category path between them. Extensions to these features have been proposed in various directions. Exploiting the ability of some machine learning algorithms to work with very large feature spaces, some authors have largely extended the representation of the constituent and its context, including among others: first and last words (and part-of-speech) in the constituent, bag-of-words, *n*-grams of part of speech, and sequence of top syntactic elements in the constituent. Parent and sibling constituents in the tree may also be codified with all the previous structural and lexical features (Pradhan et al. 2005a; Surdeanu et al. 2007). Other authors have designed new features with specific linguistic motivations. For instance, Surdeanu et al. (2003) generalized the concept of headword with the **content word** feature. They also used named entity labels as features. Xue and Palmer (2004) presented the **syntactic frame** feature, which captures the overall sentence structure using the verb predicate and the constituent as pivots. All these features resulted in a significant increase in performance.

Finally, regarding the relation between the constituent and the predicate, several variants of Gildea and Jurafsky's syntactic **path** have been proposed in the literature (e.g., generalizations to avoid sparsity, and adaptations to partial parsing). Also, some attempts have been made at characterizing the *semantic relation* between the predicate and the constituent. In Zapirain, Agirre, and Màrquez (2007) and Erk (2007), selectional preferences between predicate and headword of the constituent are explored to generate semantic compatibility features. Using conjunctions of several of the basic features is also common practice. This may be very relevant when the machine learning method used is linear in the space of features.

Joint scoring and **combination** components open the door to richer types of features, which may take into account global properties of the candidate solution plus dependencies among the different arguments. The most remarkable work in this direction is the reranking approach by Toutanova, Haghighi, and Manning in this issue. When training the ranker to select the best candidate solution they codify pattern features as strings containing the whole argument structure of the candidate. Several variations of this type of feature (with different degrees of generalization to avoid sparseness) allow them to significantly increase the performance of the base system. Also related, Pradhan et al. (2005b) and Surdeanu et al. (2007) convert the confidence scores of several base SRL systems into features for training a final machine learning–based combination system. Surdeanu et al. (2007) develop a broad spectrum of features, with sentence-based information, describing the role played by the candidate argument in every solution proposed by the different base SRL systems.

A completely different approach to feature engineering is the use of kernel methods to implicitly exploit all kinds of substructures in the syntactic representation of the candidates. This *knowledge poor* approach intends to take advantage of a massive

quantity of features without the need for manual engineering of specialized features. This motivation might be relevant for fast system development and porting, especially when specialized linguistic knowledge of the language of application is not available. The most studied approach consists of using some variants of the *'all subtrees kernel'* applied to the sentence parse trees. The work by Moschitti, Pighin, and Basili in this issue is the main representative of this family.

5. Empirical Evaluations of SRL Systems

Many experimental studies have been conducted since the work of Gildea and Jurafsky (2002), including seven international evaluation tasks in ACL-related conferences and workshops: the SIGNLL CoNLL shared tasks in 2004 and 2005 (Carreras and Màrquez 2004, 2005), the SIGLEX Senseval-3 in 2004 (Litkowski 2004), and four tasks in the SIGLEX SemEval in 2007 (Pradhan et al. 2007; Màrquez et al. 2007; Baker, Ellsworth, and Erk 2007; Litkowski and Hargraves 2007). In the subsequent sections, we summarize their main features, results, and conclusions, although note that the scores are not directly comparable across different exercises, due to differences in scoring and in the experimental methodologies.

5.1 Task Definition and Evaluation Metrics

The standard experiment in automatic SRL can be defined as follows: Given a sentence and a target predicate appearing in it, find the arguments of the predicate and label them with semantic roles. A system is evaluated in terms of precision, recall, and F_1 of the labeled arguments. In evaluating a system, an argument is considered correct when both its boundaries and the semantic role label match a gold standard. Performance can be divided into two components: (1) the precision, recall, and F_1 of unlabeled arguments, measuring the accuracy of the system at segmenting the sentence; and (2) the classification accuracy of assigning semantic roles to the arguments that have been correctly identified. In calculating the metrics, the de facto standard is to give credit only when a proposed argument perfectly matches an argument in the reference solution; nonetheless, variants that give some credit for partial matching also exist.

5.2 Shared Task Experiments Using FrameNet, PropBank, and VerbNet

To date, most experimental work has made use of English data annotated either with PropBank or FrameNet semantic roles.

The CoNLL shared tasks in 2004 and 2005 were based on PropBank (Carreras and Màrquez 2004, 2005), which is the largest evaluation benchmark available today, and also the most used by researchers—all articles in this special issue dealing with English use this benchmark. In the evaluation, the best systems obtained an F_1 score of $\sim$80%, and have achieved only minimal improvements since then. The articles in this issue by Punyakanok, Roth, and Yih; Toutanova, Haghighi, and Manning; and Pradhan, Ward, and Martin describe such efforts. An analysis of the outputs in CoNLL-2005 showed that argument identification accounts for most of the errors: a system will recall $\sim$81% of the correct unlabeled arguments, and $\sim$95% of those will be assigned the correct semantic role. The analysis also showed that systems recognized core arguments better than adjuncts (with F_1 scores from the high 60s to the high 80s for the former, but below 60% for the latter). Finally, it was also observed that, although systems performed better

on verbs appearing frequently in training, the best systems could recognize arguments of unseen verbs with an F_1 in the low 70s, not far from the overall performance.[1]

SemEval-2007 included a task on semantic evaluation for English, combining word sense disambiguation and SRL based on PropBank (Pradhan et al. 2007). Unlike the CoNLL tasks, this task concentrated on 50 selected verbs. Interestingly, the data was annotated using verb-independent roles using the PropBank/VerbNet mapping from Yi, Loper, and Palmer (2007). The two participating systems could predict VerbNet roles as accurately as PropBank verb-dependent roles.

Experiments based on FrameNet usually concentrate on a selected list of frames. In Senseval-3, 40 frames were selected for an SRL task with the goal of replicating Gildea and Jurafsky (2002) and improving on them (Litkowski 2004). Participants were evaluated on assigning semantic roles to given arguments, with best F_1 of 92%, and on the task of segmenting and labeling arguments, with best F_1 of 83%.

SemEval-2007 also included an SRL task based on FrameNet (Baker, Ellsworth, and Erk 2007). It was much more complete, realistic, and difficult than its predecessor in Senseval-3. The goal was to perform complete analysis of semantic roles on unseen texts, first determining the appropriate frames of predicates, and then determining their arguments labeled with semantic roles. It also involved creating a graph of the sentence representing part of its semantics, by means of frames and labeled arguments. The test data of this task consisted of novel manually-annotated documents, containing a number of frames and roles not in the FrameNet lexicon. Three teams submitted results, with precision percentages in the 60s, but recall percentages only in the 30s.

To our knowledge, there is no evidence to date on the relative difficulty of assigning FrameNet or PropBank roles.

5.3 Impact of Syntactic Processing in SRL

Semantic roles are closely related to syntax, and, therefore, automatic SRL heavily relies on the syntactic structure of the sentence. In PropBank, over 95% of the arguments match with a single constituent of the parse tree. If the output produced by a statistical parser is used (e.g., Collins's or Charniak's) the exact matching is still over 90%. Moreover, some simple rules can be used to join constituents and fix a considerable portion of the mismatches (Toutanova, Haghighi, and Manning 2005). Thus, it has become a common practice to use full parse trees as the main source for solving SRL.

The joint model presented in this issue by Toutanova, Haghighi, and Manning obtains an F_1 at $\sim$90% on the WSJ test of the CoNLL-2005 evaluation when using gold-standard trees; but with automatic syntactic analysis, its best result falls to $\sim$80%. This and other work consistently show that the drop in performance occurs in identifying argument boundaries; when arguments are identified correctly with predicted parses, the accuracy of assigning semantic roles is similar to that with correct parses.

A relevant question that has been addressed in experimental work concerns the use of a partial parser instead of a parser that produces full WSJ trees. In the CoNLL-2004 task, systems were restricted to the use of base syntactic phrases (i.e., *chunks*) and clauses, and the best results that could be obtained were just below 70%. But the training set in that evaluation was about five times smaller than that of the 2005 task. Punyakanok, Roth, and Yih (this issue) and Surdeanu et al. (2007) have shown that, in

1 The analysis summarized here was presented in the oral session at CoNLL-2005. The slides of the session, containing the results supporting this analysis, are available in the CoNLL-2005 shared task Web site.

fact, a system working with partial parsing can do almost as well as a system working with full parses, with differences in F_1 of only ∼2–3 points.

Currently, the top-performing systems on the CoNLL data make use of several outputs of syntactic parsers, as discussed in Section 4. It is clear that many errors in SRL are caused by having incorrect syntactic constituents, as reported by Punyakanok, Roth, and Yih in this issue. By using many parses, the recognition of semantic roles is more robust to parsing errors. Yet, it remains unanswered what is the most appropriate level of syntactic analysis needed in SRL.

5.4 Generalization of SRL Systems to New Domains

Porting a system to a new domain, different than the domain used to develop and train the system, is a challenging question in NLP. SRL is no exception, with the particular difficulty that a predicate in a new domain may exhibit a behavior not contemplated in the dictionary of frames at training time. This difficulty was identified as a major challenge in the FrameNet-based task in SemEval-2007 (Baker, Ellsworth, and Erk 2007).

In the CoNLL-2005 task, WSJ-trained systems were tested on three sections of the Brown corpus annotated by the PropBank team. The performance of all systems dropped dramatically: The best systems scored F_1 below 70%, as opposed to figures at ∼80% when testing on WSJ data. This is perhaps not surprising, taking into account that the pre-processing systems involved in the analysis (tagger and parser) also experienced a significant drop in performance. The article in this issue by Pradhan, Ward, and Martin further investigates the robustness across text genres when porting a system from WSJ to Brown. Importantly, the authors claim that the loss in accuracy takes place in assigning the semantic roles, rather than in the identification of argument boundaries.

5.5 SRL on Languages Other Than English

SemEval-2007 featured the first evaluation exercise of SRL systems for languages other than English, namely for Spanish and Catalan (Màrquez et al. 2007). The data was part of the CESS-ECE corpus, consisting of ∼100K tokens for each language. The semantic role annotations are similar to PropBank, in that role labels are specific to each verb, but also include a verb-independent thematic role label similar to the scheme proposed in VerbNet. The task consisted of assigning semantic class labels to target verbs, and identifying and labeling arguments of such verbs, in both cases using gold-standard syntax. Only two teams participated, with best results at ∼86% for disambiguating predicates, and at ∼83% for labeling arguments.

The work by Xue in this issue studies semantic role labeling for Chinese, using the Chinese PropBank and NomBank corpora. Apart from working also with nominalized predicates, this work constitutes the first comprehensive study on SRL for a language different from English.

5.6 SRL with Other Parts-of-Speech

The SemEval-2007 task on disambiguating prepositions (Litkowski and Hargraves 2007) used FrameNet sentences as the training and test data, with over 25,000 sentences for the 34 most common English prepositions. Although not overtly defined as semantic role labeling, each instance was characterized with a semantic role name and also had an associated FrameNet frame element. Almost 80% of the prepositional phrases in the instances were identified as core frame elements, and are likely to be closely associated

with arguments of the words to which they are attached. The three participants used a variety of methods, with the top performing team using machine learning techniques similar to those in other semantic role labeling tasks.

6. Final Remarks

To date, SRL systems have been shown to perform reasonably well in some controlled experiments, with F_1 measures in the low 80s on standard test collections for English. Still, a number of important challenges exist for future research on SRL. It remains unclear what is the appropriate level of syntax needed to support robust analysis of semantic roles, and to what degree improved performance in SRL is constrained by the state-of-the-art in tagging and parsing. Beyond syntax, the relation of semantic roles to other semantic knowledge (such as WordNet, named entities, or even a catalogue of frames) has scarcely been addressed in the design of current SRL models. A deeper understanding of these questions could help in developing methods that yield improved generalization, and that are less dependent on large quantities of role-annotated training data.

Indeed, the requirement of most SRL approaches for such training data, which is both difficult and highly expensive to produce, is the major obstacle to the widespread application of SRL across different genres and different languages. Given the degradation of performance when a supervised system is faced with unseen events or a testing corpus different from training, this is a major impediment to increasing the application of SRL even within English, a language for which two major annotated corpora are available. It is critical for the future of SRL that research broadens to include wider investigation of unsupervised and minimally supervised learning methods.

In addition to these open research problems, there are also methodological issues that need to be addressed regarding how research is conducted and evaluated. Shared task frameworks have been crucial in SRL development by supporting explicit comparisons of approaches, but such benchmark testing can also overly focus research efforts on small improvements in particular evaluation measures. Improving the entire SRL approach in a significant way may require more open-ended investigation and more qualitative analysis.

Acknowledgments

We are grateful for the insightful comments of two anonymous reviewers whose input helped us to improve the article. This work was supported by the Spanish Ministry of Education and Science (Màrquez); the Catalan Ministry of Innovation, Universities and Enterprise; and a grant from NTT, Agmt. Dtd. 6/21/1998 (Carreras); and NSERC of Canada (Stevenson).

References

Baker, C., M. Ellsworth, and K. Erk. 2007. SemEval-2007 Task 19: Frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic.

Boas, H. C. 2002. Bilingual framenet dictionaries for machine translation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1364–1371, Las Palmas de Gran Canaria, Spain.

Briscoe, T. and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing (ANLP)*, pages 356–363, Washington, DC.

Carreras, X. and L. Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, MA.

Carreras, X. and L. Màrquez. 2005. Introduction to the CoNLL-2005 Shared

Task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, MI.

Carreras, X., L. Màrquez, and G. Chrupała. 2004. Hierarchical recognition of propositional arguments with perceptrons. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 106–109, Boston, MA.

Cohn, T. and P. Blunsom. 2005. Semantic role labelling with tree conditional random fields. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 169–172, Ann Arbor, MI.

Copestake, A. and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 591–600, Athens, Greece.

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.

Erk, K. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic.

Fillmore, C. 1968. The case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart & Winston, New York, pages 1–88.

Fillmore, C. J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20–32.

Fillmore, C. J., J. Ruppenhofer, and C. F. Baker. 2004. Framenet and representing the link between semantic and syntactic relations. In Churen Huang and Winfried Lenders, editors, *Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B*. Institute of Linguistics, Academia Sinica, Taipei, pages 19–59.

Gildea, D. and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Grimshaw, J. 1990. *Argument Structure*. MIT Press, Cambridge, MA.

Habash, N., B. J. Dorr, and D. Traum. 2003. Hybrid natural language generation from lexical conceptual structures. *Machine Translation*, 18(2):81–128.

Hirst, G. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, Cambridge.

Jackendoff, R. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

Johansson, R. and P. Nugues. 2007. LTH: Semantic structure extraction using nonprojective dependency trees. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230, Prague, Czech Republic.

Kipper, K., H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX.

Koomen, P., V. Punyakanok, D. Roth, and W. Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 181–184, Ann Arbor, MI.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago, IL.

Levin, B. and M. Rappaport Hovav. 1998. Building verb meanings. In M. Butt and W. Geuder, editors, *The Projection of Arguments: Lexical and Compositional Factors*. CSLI Publications, Stanford, CA, pages 97–134.

Levin, B. and M. Rappaport Hovav. 2005. *Argument Realization*. Cambridge University Press, Cambridge.

Litkowski, K. C. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 9–12, Barcelona, Spain.

Litkowski, K. C. and O. Hargraves. 2005. The preposition project. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*, pages 171–179, Colchester, UK.

Litkowski, K. C. and O. Hargraves. 2007. SemEval-2007 Task 06: Word-sense disambiguation of prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic.

Loper, E., S. Yi, and M. Palmer. 2007. Combining lexical resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 118–128, Tilburg, The Netherlands.

Màrquez, L., P. R. Comas, J. Giménez, and N. Català. 2005. Semantic role labeling as sequential tagging. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 193–196, Ann Arbor, MI.

Màrquez, L., L. Villarejo, M.A. Martí, and M. Taulé. 2007. SemEval-2007 Task 09: Multilevel semantic annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 42–47, Prague, Czech Republic.

Melli, G., Y. Wang, Y. Liu, M. M. Kashani, Z. Shi, B. Gu, A. Sarkar, and F. 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.

Merlo, P. and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An interim report. In *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, MA.

Moldovan, D., A. Badulescu, M. Tatu, D. Antohe, and R. Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL 2004 Workshop on Computational Lexical Semantics*, pages 60–67, Boston, MA.

Musillo, G. and P. Merlo. 2006. Accurate parsing of the proposition bank. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 101–104, New York, NY.

Narayanan, S. and S. Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 693–701, Geneva, Switzerland.

O'Hara, T. and J. Wiebe. 2003. Preposition semantic classification via Penn Treebank and FrameNet. In *Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, pages 79–86, Edmonton, Canada.

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pradhan, S., K. Hacioglu, V. Krugler, W. Ward, J. Martin, and D. Jurafsky. 2005a. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Pradhan, S., K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky. 2005b. Semantic role chunking combining complementary syntactic views. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 217–220, Ann Arbor, MI.

Pradhan, S., E. Loper, D. Dligach, and M. Palmer. 2007. SemEval-2007 Task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Rosario, B. and M. Hearst. 2004. Classifying semantic relations in bioscience text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 430–437, Barcelona, Spain.

Schulte im Walde, S. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Shi, L. and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing; Sixth International Conference, CICLing 2005, Proceedings*, LNCS, vol 3406, pages 100–111, Mexico City, Mexico.

Surdeanu, M., S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.

Surdeanu, M., L. Màrquez, X. Carreras, and P. R. Comas. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research (JAIR)*, 29:105–151.

Swier, R. and S. Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 95–102, Barcelona, Spain.

Swier, R. and S. Stevenson. 2005. Exploiting a verb lexicon in automatic semantic role

labelling. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMLNP)*, pages 883–890, Vancouver, B.C., Canada.

Thompson, C. A., R. Levy, and C. Manning. 2003. A generative model for semantic role labeling. In *Proceedings of the 14th European Conference on Machine Learning (ECML)*, pages 397–408, Dubrovnik, Croatia.

Toutanova, K., A. Haghighi, and C. Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 589–596, Ann Arbor, MI.

Xue, N. and M. Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 88–94, Barcelona, Spain.

Yi, S., E. Loper, and M. Palmer. 2007. Can semantic roles generalize across corpora? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 548–555, Rochester, NY.

Zapirain, B., E. Agirre, and L. Màrquez. 2007. UBC-UPC: Sequential SRL using selectional preferences: an approach with maximum entropy Markov models. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 354–357, Prague, Czech Republic.