Extracting a verb lexicon for deep parsing from FrameNet

Mark McConville and **Myroslava O. Dzikovska**
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland
{Mark.McConville,M.Dzikovska}@ed.ac.uk

Abstract

We examine the feasibility of harvesting a wide-coverage lexicon of English verbs from the FrameNet semantically annotated corpus, intended for use in a practical natural language understanding (NLU) system. We identify a range of constructions for which current annotation practice leads to problems in deriving appropriate lexical entries, for example imperatives, passives and control, and discuss potential solutions.

1 Introduction

Although the lexicon is the primary source of information in lexicalised formalisms such as HPSG or CCG, constructing one manually is a highly labour-intensive task. Syntactic lexicons *have* been derived from other resources — the LinGO ERG lexicon (Copestake and Flickinger, 2000) contains entries extracted from ComLex (Grishman et al., 1994), and Hockenmaier and Steedman (2002) acquire a CCG lexicon from the Penn Treebank. However, one thing these resources lack is information on how the syntactic subcategorisation frames correspond to meaning.

The output representation of many "deep" wide coverage parsers is therefore limited with respect to argument structure — sense distinctions are strictly determined by syntactic generalisations, and are not always consistent. For example, in the logical form produced by the LinGO ERG grammar, the verb *end* can have one of two senses depending on its subcategorisation frame: `end_v_1_rel` or `end_v_cause_rel`, corresponding to *the celebrations ended* and *the storm ended the celebrations* respectively. Yet a very similar verb, *stop*, has a single sense, `stop_v_1_rel`, for both *the celebrations stopped* and *the storm stopped the celebrations*. There is no direct connection between these different verbs in the ERG lexicon, even though they are intuitively related and are listed as belonging to the same or related word classes in semantic lexicons/ontologies such as VerbNet (Kipper et al., 2000) and FrameNet (Baker et al., 1998).

If the output of a deep parser is to be used with a knowledge representation and reasoning component, for example in a dialogue system, then we need a more consistent set of word senses, linked by specified semantic relations. In this paper, we investigate how straightforward it is to harvest a computational lexicon containing this kind of information from FrameNet, a semantically annotated corpus of English. In addition, we consider how the FrameNet annotation system could be made more transparent for lexical harvesting.

Section 2 introduces the FrameNet corpus, and section 3 discusses the lexical information required by frame-based NLU systems, with particular emphasis on linking syntactic and semantic structure. Section 4 presents the algorithm which converts the FrameNet corpus into a frame-based lexicon, and evaluates the kinds of entries harvested in this way. We then discuss a number of sets of entries which are inappropriate for inclusion in a frame-based lexicon: (a) 'subjectless' entries; (b) entries derived from passive verbs; (c) entries subcategorising for modifiers; and (d) entries involving 'control' verbs.

Proceedings of the 5th Workshop on Important Unresolved Matters, pages 112–119,
Ann Arbor, June 2005.

2 FrameNet

FrameNet[1] is a corpus of English sentences annotated with both syntactic and semantic information. Underlying the corpus is an ontology of 795 'frames' (or semantic *types*), each of which is associated with a set of 'frame elements' (or semantic *roles*). To take a simple example, the `Apply_heat` frame describes a situation involving frame elements such as a `COOK`, some `FOOD`, and a `HEATING_INSTRUMENT`. Each frame is, in addition, associated with a set of 'lexical units' which are understood as *evoking* it. For example, the `Apply_heat` frame is evoked by such verbs as *bake*, *blanch*, *boil*, *broil*, *brown*, *simmer*, *steam*, etc.

The FrameNet corpus proper consists of 139,439 sentences (mainly drawn from the British National Corpus), each of which has been hand-annotated with respect to a particular target word in the sentence. Take the following example: *Matilde fried the catfish in a heavy iron skillet*. The process of annotating this sentence runs as follows: (a) identify a target word for the annotation, for example the main verb *fried*; (b) identify the semantic frame which is evoked by the target word in this particular sentence – in this case the relevant frame is `Apply_heat`; (c) identify the sentential constituents which realise each frame element associated with the frame, i.e.:

[`COOK` *Matilde*] [`Apply_heat` ***fried***] [`FOOD` *the catfish*] [`HEATING_INSTR` *in a heavy iron skillet*]

Finally, some basic syntactic information about the target word and the constituents realising the various frame elements is also added: (a) the part-of-speech of the target word (e.g. `V`, `N`, `A`, `PREP`); (b) the syntactic *category* of each constituent realising a frame element (e.g. `NP`, `PP`, `VPto`, `Sfin`); and (c) the syntactic *role*, with respect to the target word, of each constituent realising a frame element (e.g. `Ext`, `Obj`, `Dep`). Thus, each sentence in the corpus can be seen to be annotated on at least three independent 'layers', as exemplified in Figure 1.

3 Frame-based NLU

The core of any frame-based NLU system is a parser which produces domain-independent semantic representations like the following, for the sentence *John billed the champagne to my account*:

$$\begin{bmatrix} \textit{commerce-pay} & \\ \text{AGENT} & John \\ \text{THEME} & champagne \\ \text{SOURCE} & \begin{bmatrix} \textit{account} & \\ \text{OWNER} & me \end{bmatrix} \end{bmatrix}$$

Deep parsers/grammars such as the ERG, OpenCCG (White, 2006) and TRIPS (Dzikovska, 2004) produce more sophisticated representations with scoping and referential information, but still contain a frame-based representation as their core. The lexical entries necessary for constructing such representations specify information about orthography, part-of-speech, semantic type and subcategorisation properties, including a mapping between a syntactic subcategorisation frame and the semantic frame.

An example of a TRIPS lexical entry is presented in Figure 2, representing the entry for the verb *bill* as used in the sentence discussed above. Note that for each subcategorised argument the syntactic role, syntactic category, and semantic role are specified. Much the same kind of information is included in ERG and OpenCCG lexical entries.

When constructing a computational lexicon, there are a number of issues to take into account, several of which are pertinent to the following discussion. Firstly, computational lexicons typically list only the 'canonical' subcategorisation frames, corresponding to a declarative sentence whose main verb is in the active voice, as in Figure 1. Other variations, such as passive forms, imperatives and dative alternations are generated automatically, for example by lexical rules. Secondly, parsers that build semantic representations typically make a distinction between 'complements' and 'modifiers'. Complements are those dependents whose meaning is completely determined by the verb, for example the PP *on him* in the sentence *Mary relied on him*, and are thus listed in lexical entries. Modifiers, on the other hand, are generally not specified in verb entries — although they may be associated with the underlying verb frame, their meaning is determined independently, usually by the preposition, such as the time adverbial *next week* in *I will see him next week*.

Finally, for deep parsers, knowledge about which argument of a matrix verb 'controls' the implicit

[1]The version of FrameNet discussed in this paper is FrameNet II release 1.3 from 22 August 2006.

	Matilde	*fried*	*the catfish*	*in a heavy iron skillet*
target		Apply_heat		
frame element	COOK		FOOD	HEATING_INSTR
syntactic category	NP	V	NP	PP
syntactic role	Ext		Obj	Dep

Figure 1: A FrameNet annotated sentence

$$\begin{bmatrix} \text{ORTH} & \langle \textit{bill} \rangle \\ \text{SYNCAT} & \textit{v} \\ \text{SEMTYPE} & \begin{bmatrix} \textit{commerce-pay} & \\ \text{ASPECT} & bounded \\ \text{TIME-SPAN} & atomic \end{bmatrix} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{SYNROLE} & subj \\ \text{SYNCAT} & np \\ \text{SEMROLE} & agent \end{bmatrix}, \begin{bmatrix} \text{SYNROLE} & obj \\ \text{SYNCAT} & np \\ \text{SEMROLE} & theme \end{bmatrix}, \begin{bmatrix} \text{SYNROLE} & comp \\ \text{SYNCAT} & \begin{bmatrix} \textit{pp} & \\ \text{PTYPE} & to \end{bmatrix} \\ \text{SEMROLE} & source \end{bmatrix} \right\rangle \end{bmatrix}$$

Figure 2: A TRIPS lexical entry

subject of an embedded complement verb phrase is necessary in order to to build the correct semantic form. In a unification parser such as TRIPS, control is usually represented by a relation of token-identity (i.e. feature structure reentrancy) between the subject or object of a control verb and the subject of a verbal complement.

4 Harvesting a computational lexicon from FrameNet

In order to harvest a computational lexicon from the FrameNet corpus, we took each of the 60,309 annotated sentences whose target word is a verb and derived a lexical entry directly from the annotated information. For example, from the sentence in Figure 1, the lexical entry in Figure 3 is derived.[2]

In order to remove duplicate entries, we made two assumptions: (a) the value of the ARGS feature is a *set* of arguments, rather than, say, a list or multiset; and (b) two arguments are identical just in case they specify the same syntactic role and semantic role. These assumptions prevent a range of inappropriate entries from being created, for example entries derived from sentences involving a 'split' argument, both parts of which are annotated independently in FrameNet, e.g. [Ext *Serious concern*] ***arose*** [Ext *about his motives*]. A second group of inappropriate entries which are thus avoided are those deriving from relative clause constructions, where the relative pronoun and its antecedent are also annotated separately:

[Ext Perp *The two boys*] [Ext Perp *who*] ***abducted*** [Obj Victim *James Bulger*] *are likely to have been his murderers*

Finally, assuming that the arguments constitute a set means that entries derived from sentences involving both canonical[3] and non-canonical word order are treated as equivalent. The kinds of construction implicated here include 'quotative inversion' (e.g. *"Or Electric Ladyland," added Bob*), and leftwards extraction of objects and dependents, for example:

Are there [Obj *any places*] [Ext *you*] *want to* ***praise*** [Dep *for their special facilities*]?

In this paper we are mainly interested in extracting the possible syntax-semantics mappings from FrameNet, rather than the precise details of their relative ordering. Since dependents in the harvested

[2]Our original plan was to use the automatically generated 'lexical entry' files included with the most recent FrameNet release as a basis for deep parsing. However, these entries contain so many inappropriate subcategorisation frames, of the types discussed in this paper, that we decided to start from scratch with the corpus annotations.

[3]The canonical word order in English involves a pre-verbal subject, with all other dependents following the verb.

$$\begin{bmatrix} \text{ORTH} & \langle \textit{fry} \rangle \\ \text{SYNCAT} & \texttt{V} \\ \text{SEMTYPE} & \texttt{Apply\_heat} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{SYNROLE} & \texttt{Ext} \\ \text{SYNCAT} & \texttt{NP} \\ \text{SEMROLE} & \texttt{Cook} \end{bmatrix} \begin{bmatrix} \text{SYNROLE} & \texttt{Obj} \\ \text{SYNCAT} & \texttt{NP} \\ \text{SEMROLE} & \texttt{Food} \end{bmatrix}, \begin{bmatrix} \text{SYNROLE} & \texttt{Dep} \\ \text{SYNCAT} & \texttt{PP} \\ \text{SEMROLE} & \texttt{Heating\_Instr} \end{bmatrix} \right\rangle \end{bmatrix}$$

Figure 3: The lexical entry derived from Figure 1

lexicon are fully specified for semantic role, syntactic category *and* syntactic role, post-verbal constituent ordering can-be regulated extra-lexically by means of precedence rules. For example, for the TRIPS and LFG formalisms, there is a straightforward correspondence between their native syntactic role specifications and the FrameNet syntactic roles.

After duplicate entries were removed from the resulting lexicon, we were left with 26,022 distinct entries. The harvested lexicon incorporated 2,002 distinct orthographic forms, 358 distinct frames, and 2,661 distinct orthography-frame pairs, giving a functionality ratio (average number of lexical entries per orthography-type pair) of 9.8.

Next, we evaluated a random sample of the derived lexical entries by hand. The aim here was to identify general classes of the harvested verb entries which are not appropriate for inclusion in a frame-based verb lexicon, and which would need to be identified and fixed in some way. The main groups identified were: (a) entries with no `Ext` argument (section 4.1); (b) entries derived from verbs in the passive voice (section 4.2); (c) entries which subcategorise for modifiers (section 4.3); and (d) entries for control verbs (section 4.4).

4.1 Subjectless entries

The harvested lexicon contains 2,201 entries (i.e. 9% of the total) which were derived from sentences which do *not* contain an argument labelled with the `Ext` syntactic role, in contravention of the generally accepted constraint on English verbs that they always have a subject.

Three main groups of sentences are implicated here: (a) those featuring *imperative* uses of the target verb, e.g. *Always* ***moisturise*** *exposed skin with an effective emollient like E45*; (b) those featuring other *non-finite* forms of the target verb whose understood subject is not controlled by (or even coreferential with) some other constituent in the sentence, e.g. *Being* ***accused*** *of not having a sense of humour is a terrible insult*; and (c) those involving a non-referential subject *it*, for example *It is* ***raining*** *heavily* or *It is to be* ***regretted*** *that the owner should have cut down the trees*. In FrameNet annotations, non-referential subjects are not identified on the syntactic role annotation layer, and this makes it more difficult to harvest appropriate lexical entries for these verbs from the corpus.

These entries are easy to locate in the harvested lexicon, but more difficult to repair. Typically one would want to discard the entries generated from (a) and (b) as they will be derived automatically in the grammar, but keep the entries generated from (c) while adding a non-referential *it* as a subject.

Although the FrameNet policy is to annotate the (a) and (b) sentences as having a 'non-overt' realisation of the relevant frame element, this is confined to the frame element annotation layer itself, with the syntactic role and syntactic category layers containing *no* clues whatsoever about understood subjects. One rather roundabout way of differentiating between these cases would involve attempting to identify the syntactic category and semantic role of the missing `Ext` argument by looking at other entries with the same orthography and semantic type. However, this whole problem could be avoided if understood and expletive subjects were identified on the *syntactic* layers in FrameNet annotations.

4.2 'Passive' entries

Many entries in the harvested lexicon were derived from sentences where the target verb is used in the passive voice, for example:

[`Ext NP Victim` *The men*] *had allegedly been* ***abducted*** [`Dep PP Perp` *by Mrs Mandela's body-*

guards] [Dep PP Time *in 1988*]

As discussed above, computational lexicons do not usually list the kinds of lexical entry derived directly from such sentences. Thus, it is necessary to identify and correct or remove them.

In FrameNet annotated sentences, the voice of target verbs is not marked explicitly.[4] We applied the following simple diagnostic to identify 'passive' entries: (a) there is an Ext argument realising frame element e; and (b) there is some other entry with the same orthographic form and semantic frame, which has an Obj argument realising frame element e.

Initially we applied this diagnostic to the entries in the harvested lexicon together with a part-of-speech tag filter. The current FrameNet release includes standard POS-tag information for each word in each annotated sentence. We considered only those lexical entries derived from sentences whose target verb is tagged as a 'past-participle' form (i.e. VVN). This technique identified 4,160 entries in the harvested lexicon (i.e. 16% of the total) as being 'passive'. A random sample of 10% of these was examined and *no* false positives were found.

The diagnostic test was then repeated on the remaining lexical entries, this time *without* the POS-tag filter. This was deemed necessary in order to pick up false negatives caused by the POS-tagger having assigned the wrong tag to a passive target verb (generally the past tense form tag VVD). This test identified a further 1007 entries as 'passive' (4% of the total entries). As well as mis-tagged instances of normal passives, this test picked up a further three classes of entry derived from target verbs appearing in passive-related constructions. The first of these involves cases where the target verb is in the complement of a 'raising adjective' (e.g. *tough*, *difficult*, *easy*, *impossible*), for example:

[Ext NP Goal *Both planning and control*] *are difficult to* ***achieve*** [Dep PP Circs *in this form of production*]

The current FrameNet annotation guidelines (Ruppenhofer et al., 2006) state that the extracted object in these cases *should* be tagged as Obj. However, in practice, the majority of these instances appear to have been tagged as Ext.

[4]Whilst there are dedicated subcorpora containing *only* passive targets, it is not the case that *all* passive targets are in these.

The second group of passive-related entries involve verbs in the *need -ing* construction[5], e.g.:

[Ext NP Content *Many private problems*] *need* ***airing*** [Dep PP Medium *in the family*]

The third group involved sentences where the target verb is used in the 'middle' construction:

[Ext Experiencer *You*] ***frighten*** [Dep Manner *easily*]

Again, linguistically-motivated grammars generally treat these three constructions in the rule component rather than the lexicon. Thus, the lexical entries derived from these sentences need to be located and repaired, perhaps by comparison with other entries.

Of the 1007 lexical entries identified by the second, weaker form of the passive test, 224 (i.e. 22%) turn out to be false positives. The vast majority of these involve verbs implicated in the causative-inchoative alternation (e.g. *John's back arched* vs. *John arched his back*). The official FrameNet policy is to distinguish between frames encoding a change-of-state and those encoding the causation of such a change, for example Amalgamation versus Cause_to_amalgamate, Motion versus Cause_motion etc. In each case, the two frames are linked by the Causative_of frame relation. Most of the false positives are the result of a failure to consistently apply this principle in annotation practice, for example where no causative counterpart has been defined for a particular inchoative frame, or where an inchoative target has been assigned to a causative frame, or a causative target to an inchoative frame. For example, 94 of the false positives are accounted for simply by the lack of a causative counterpart for the Body_movement frame, meaning that both inchoative and causative uses of verbs like *arch*, *flutter* and *wiggle* have all been assigned to the same frame.

For reasons of data sparsity, it is expected that the approach to identifying passive uses of target verbs discussed here will result in false negatives, since it relies on there being at least one corresponding active use in the corpus. We checked a random sample of 400 of the remaining entries in the harvested lexicon and found nine false negatives, suggesting that

[5]Alternatively *merit -ing*, *bear -ing* etc.

the test successfully identifies 91% of those lexical entries derived from passive uses of target verbs.

4.3 Modifiers

General linguistic theory makes a distinction between two kinds of non-subject dependent of a verb, depending on the notional ‘closeness’ of the semantic relation — complements vs. modifiers. Take for example the following sentence:

[Ext Performer *She*]*'s* [Dep Time *currently*] ***starring*** [Dep Performance *in The Cemetery Club*] [Dep Place *at the Wyvern Theatre*]

Of the three constituents annotated here as Dep, only one is an complement (the Performance); the Time and Place dependents are modifiers. Frame-based NLU systems do not generally list modifiers in the argument structure of a verb’s lexical entry. Thus, we need to find a means of identifying those dependents in the harvested lexicon which are actually modifiers.

The FrameNet ontology provides some information to help differentiate complements and modifiers. A frame element can be marked as Core, signifying that it “instantiates a conceptually necessary component of a frame, while making the frame unique and different from other frames”. The annotation guidelines state that the distinction between Core and non-Core frame elements covers “the semantic spirit” of the distinction between complements and modifiers. Thus, for example, obligatory dependents are always Core, as are: (a) those which, when omitted, receive a definite interpretation (e.g. the Goal in *John arrived*); and (b) those whose semantics cannot be predicted from their form. In the Performers_and_roles frame used in the example above, the Performer and Performance frame elements are marked as Core, whilst Time and Place are not.

However, it is not clear that the notion of ontological ‘coreness’ used in FrameNet corresponds well with the intuitive distinction between syntactic complements and modifiers. This is exemplified by the existence of numerous constituents in the corpus which have been marked as direct objects, despite invoking non-*Core* frame elements, for example:

[Agent *I*] **ripped** [Subregion *the top*] [Patient *from my packet of cigarettes*]

The relevant frame here is Damaging, where the Subregion frame element is marked as non-Core, based on examples like *John ripped his trousers* [*below the knee*]. Thus in this case, the decision to retain all senses of the verb *rip* within the same frame has led to a situation where semantic and syntactic coreness have become dislocated. Thus, although the Core vs. non-Core property on frame elements *does* yield a certain amount of information about which arguments are complements and which are modifiers, greater care needs to be taken when assigning different subcategorisation alternants to the same frame. For example, it would have been more convenient to have assigned the verb *rip* in the above example to the Removing frame, where the direct object would then be assigned the Core frame element Theme.

In the example discussed above, FrameNet does provide syntactic role information (Obj) allowing us to infer that a non-Core role is a complement rather than a modifier. Where the syntactic role is simply marked as Dep however, it is not possible to make the decision without recourse to other lexical resources (e.g. ComLex). Since different parsers may utilise different criteria for distinguishing complements from modifiers, it might be better to postpone this task to the syntactic alignment module.

4.4 Control verbs

Unification-based parsers generally handle the distinction between subject (*John promised Mary to go*) and object (*John persuaded Mary to go*) control verbs in the lexicon, using coindexation of the subject/object of the control verb and the understood subject of the embedded verb. The parser can use this lexical information to assign the correct referent to the understood subject in a sentence like *John asked Mary to go*:

$$\begin{bmatrix} \textit{command} & \\ \text{AGENT} & John \\ \text{THEME} & Mary\boxed{1} \\ \text{EFFECT} & \begin{bmatrix} \textit{motion} & \\ \text{THEME} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

Control verbs are annotated in FrameNet in the following manner:

Perhaps [Ext NP Speaker *we*] *can* ***persuade*** [Obj NP Addressee *Tammuz*] [Dep VPto

`Content` *to entertain him*]

The lexical entries for transitive control verbs that we can harvest directly from these annotations thus fail to identify whether it is the subject or the direct object which controls the understood subject of the embedded verb.

We attempted to automatically distinguish subject from object control in FrameNet by looking for the annotated sentences that contain independently annotated argument structures for both the control verb and embedded verb. For example, let's assume the following annotation also exists in the corpus:

Perhaps we can persuade [`Ext NP Agent` *Tammuz*] *to* ***entertain*** [`Obj NP Experiencer` *him*]

We can then use the fact that it is the *object* of the control verb which is coextensive with the `Ext` of the embedded verb to successfully identify *persuade* as an object-control verb.

The problem with this approach is data sparsity. The harvested lexicon contains 135 distinct verbs which subcategorise for both a direct object and a controlled VP complement. In a random sample of ten of these *none* of the annotated sentences had been annotated independently from the perspective of the governed verb. As the proportion of the FrameNet corpus which involves annotation of running text, rather than cherry-picked example sentences, increases, we would expect this to improve.[6]

5 Implementation and discussion

The revised version of the harvested lexicon contains 9,019 entries for 2,626 orthography-frame pairs, yielding a functionality ratio of 3.4.

This lexicon still requires a certain amount of cleaning up. For example, the verb *accompany* is assigned to a number of distinct lexical entries depending on the semantic role associated with the PP complement (i.e. `Goal`, `Path` or `Source`). Cases like this, where the role name is determined by the particular choice of preposition, could be handled outside the lexicon. Alternatively, it may be possible to use the 'core set' feature of the FrameNet ontology (which groups together roles that are judged to be equivalent in some sense) to locate this kind of redundancy. Other problems involve sentences where a possessive determiner has been annotated as the subject of a verb, e.g. *It was [his] intention to* ***aid*** *Larsen*, resulting in numerous spurious entries.

The harvested lexical entries are encoded according to a framework-independent XML schema, which we developed with the aim of deriving lexicons for use with a diverse range of parsers. At the moment, several additional steps are required to convert the entries we extracted into a format suitable for a particular parser.

Firstly, the syntactic categories used by FrameNet and the target lexicon have to be reconciled. While basic constituent types such as noun and adjective phrases do not change between the theories, small differences may still exist. For example, the TRIPS parser classifies all *wh*-clauses such as *what he did* in *I saw what he did* and *What he did was good* as noun phrases, the LinGO ERG grammar interprets them as either noun phrases or clauses depending on the context, and FrameNet annotation classifies all of them as clauses. The alignment, however, should be relative straightforward as there is, in general, good agreement on the basic syntactic categories.[7]

Secondly, the information relevant to constituent ordering may need to be derived, as discussed in Section 4. Finally, the more abstract features such as control have to be converted into feature structures appropriate for the unification parsers. Our schema incorporates the possibility for embedded category structure, as in the treatment of control verbs in CCG and HPSG where the verbal complement is an 'unsaturated' category. We plan to use our schema as a platform for deriving richer lexical representations from the 'flatter' entries harvested directly from FrameNet.

As part of our future work, we expect to create generic algorithms that help automate these steps. In particular, we plan to include a domain-independent set of constituent categories and syntactic role labels, and add algorithms that convert between a linear ordering and a set of functional labels, for example (Crabbé et al., 2006). We also plan to develop algorithms to import information from other seman-

[6]An alternative approach would be to consult an external lexical resource, e.g. the LinGO ERG lexicon, which has good coverage of control verbs.

[7]http://www.cl.cam.ac.uk/users/alk23/classes/Classes2.txt contains a list of mappings between three different deep parsers and ComLex subcategorisation frames

tic lexicons such as VerbNet into the same schema.

Currently, we have implemented an algorithm for converting the harvested entries into the TRIPS lexicon format, resulting in a 6133 entry verb lexicon involving 2654 distinct orthography-type pairs. This lexicon has been successfully used with the TRIPS parser, but additional work remains to be done before the conversion process is complete. For example, we need a more sophisticated approach to resolving the complement-modifier distinction, along with a means of integrating the FrameNet semantic types with the TRIPS ontology so the parser can use selectional restrictions to disambiguate.

The discussion in this paper has been mainly focused on extracting entries for a deep lexicons using frame-based NLU, but similar issues have been faced also by the developers of shallow semantic parsers from semantically annotated corpora. For example, Gildea and Jurafsky (2002) found that identifying passives was important in training a semantic role classifier from FrameNet, using a parser trained on the Penn Treebank along with a set of templates to distinguish passive constructions from active ones. Similarly, Chen and Rambow (2003) argue that the kind of deep linguistic features we harvest from FrameNet is beneficial for the successful assignment of PropBank roles to constituents, in this case using TAGs generated from PropBank to generate the relevant features. From this perspective, our harvested lexicon can be seen as providing a 'cleaned-up', filtered version of FrameNet for training semantic interpreters. It may also be utilised to provide information for a separate lexical interpretation and disambiguation module to be built on top of a syntactic parser.

6 Conclusion

We have developed both a procedure and a framework-independent representation schema for harvesting lexical information for deep NLP systems from the FrameNet semantically annotated corpus. In examining the feasibility of this approach to increasing lexical coverage, we have identified a number of constructions for which current FrameNet annotation practice leads to problems in deriving appropriate lexical entries, for example imperatives, passives and control.

7 Acknowledgements

The work reported here was supported by grants N000140510043 and N000140510048 from the Office of Naval Research.

References

C. F. Baker, C. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98, Montreal*, pages 86–90.

J. Chen and O. Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In *Proceedings of EMNLP'03, Sapporo, Japan*.

A. Copestake and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of LREC'00, Athens, Greece*, pages 591–600.

B. Crabbé, M. O. Dzikovska, W. de Beaumont, and M. Swift. 2006. Increasing the coverage of a domain independent dialogue lexicon with VerbNet. In *Proceedings of ScaNaLU'06*, New York City.

M. O. Dzikovska. 2004. *A Practical Semantic Representation for Natural Language Parsing*. Ph.D. thesis, University of Rochester, Rochester NY.

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

R. Grishman, C. MacLeod, and A. Meyers. 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of COLING'94, Kyoto, Japan*, pages 268–272.

J. Hockenmaier and M. Steedman. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. In *Proceedings of LREC'02, Las Palmas, Spain*.

K. Kipper, H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of AAAI'00, Austin TX*.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk, 2006. *FrameNet II: Extended Theory and Practice*. The Berkeley FrameNet Project, August.

M. White. 2006. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.