Utilization of Multiple Language Resources for Robust Grammar-Based Tense and Aspect Classification

Alexis Palmer, Jonas Kuhn, Carlota Smith

The University of Texas at Austin
Department of Linguistics
Austin, TX 78712, USA
EMAIL {alexispalmer,jonask,carlotasmith}@mail.utexas.edu

Abstract

This paper reports on an ongoing project that uses varied language resources and advanced NLP tools for a linguistic classification task in discourse semantics. The system we present is designed to assign a "situation entity" class label to each predicator in English text. The project goal is to achieve the best-possible identification of situation entities in naturally-occurring written texts by implementing a robust system that will deal with real corpus material, rather than just with constructed textbook examples of discourse. In this paper we focus on the combination of multiple information sources, which we see as being vital for a robust classification system. We use a deep syntactic grammar of English to identify morphological, syntactic, and discourse clues, and we use various lexical databases for fine-grained semantic properties of the predicators. Experiments performed to date show that enhancing the output of the grammar with information from lexical resources improves recall but lowers precision in the situation entity classification task.

1. Introduction

1.1. Situation entities and the discourse semantic theory

The system we present is designed to assign a "situation entity" class label to each predicator (verb, nominal) in written English text. The broader context of this work is the theory of Modes of Discourse (Smith, 2003), which classifies discourse passages into discourse modes such as Narrative, Description, Argument, Information, and Report. The discourse modes are distinguished by the status of the text passage with respect to two factors: the situation entities in the passage and their temporal relation. The present work addresses only the first of those factors.

Situation entities in (Smith, 2003) receive a three-way categorization, with further distinctions made within each of the three categories. This paper appeals only to the three general categories: eventualities (particular events and states), generalizing statives (generics and habitual states), and abstract entities (embedded facts and propositions). The discourse modes characteristically introduce different types of situation entities. Text passages in the Narrative and Report modes generally introduce events and states, for example, while the Argument mode primarily introduces abstract entities (facts and propositions) and general statives.

Consider the following text fragment as a typical Narrative mode text passage.[1]

(1) a. One rainy afternoon she was at the center sitting with the boy.
b. He had been spinning a puzzle piece for the past twenty minutes.
c. My mom had picked up another puzzle piece and was spinning it too,
d. smiling encouragingly at him to show him how much fun they were having together.
e. Without warning, the boy dropped his puzzle piece in my mother's lap,
f. stood up, and
g. walked a few feet towards the wall of the room.

The situation entities introduced by this passage are four states (1a,b,c,d) and three events (1e,f,g). The system we present in this paper is designed to recognize and classify situation entities.

1.2. Linguistic correlates of situation entities

Situation entities are identifiable using a number of linguistic tests (again see (Smith, 2003) for details).

Automatic aspectual classification is problematic because such classification relies on a broad range of linguistic information (lexical, contextual, syntactic), and these various levels of information are not often all present within a single system. In addition, the tests for aspectual classification are generally not robust enough to allow for satisfactory results in automatic classification tasks, in particular when classification is occurring on real text examples. Our approach relies on the ordered combination of clues from various levels of representation. We strengthen the results of weaker linguistic tests by applying them only after stronger, more specific tests have failed. Any approach (ours included) of course is impacted by the fact that in real text only a subset of the theoretically available linguistic clues will be present and automatically identifiable.

The primary resource of the system is the broad-coverage English LFG-grammar developed by PARC

[1] Taken from the website Other People's Stories. http://www.otherpeoplesstories.com (Gentile, Andi. "On a boy from another dimension.")

(the Palo Alto Research Center) in the context of the Parallel Grammar Development project (ParGram, http://www2.parc.com/istl/groups/nltt/pargram/). Using PARC's XLE parsing system, each sentence of the input text is parsed and assigned a forest of phrase structure trees and feature structure (f-structure) representations by the grammar. To pick a particular reading, we apply the statistical disambiguation component of (Riezler et al., 2002). The output of XLE for a sentence is a deep predicate-argument structure analysis including morphosyntactic features. A robustness component provides partial subconstituent analyses for sentences failing to receive a full parse. Thus every input string is parsed, and linguistic tests are applied to the resulting f-structures.

For example, a well-known test for generic constructions is the presence of a bare plural subject (following (Carlson and Pelletier, 1995), among others), as in (2).

(2) Lions sleep in the shade.

The f-structure representation output (shown in Figure 1) by the XLE parser contains all of the linguistic features (number, lack of specifier, subject position) needed to identify the predicate 'lions' as the subject of a generic statement in this sentence. Our system looks into the f-structure representation to identify these features and thereby classify the sentence as a general stative.

Three two-valued temporal features hold of events and states: dynamic-static, telic-atelic, durative-instantaneous. This temporal/aspectual information also plays a role in the identification of situation entities, yet it is for the most part not represented in the output of the parser. To get this information we appeal to various lexical resources, as described in Section 2.1.

2. Current approach and system architecture

Our approach is to appeal to a number of sources of information, using ordering of tests to weight the various sources of information and arrive at the most-likely classification for each situation entity. In bringing together the various information sources we use an existing NLP tool that was originally developed at PARC for transfer in Machine Translation. This tool allows us to transform a rich syntactic analysis into a situation entity representation by exploiting clues from various levels of representation.

At a high level, our system has three components: parsing of the sentences of a text using the English LFG-grammar, augmentation of the grammar output representations, and application of linguistic tests.

Input texts are preprocessed into XLE testfile format, and each testfile is run through the XLE parser, using the statistical disambiguation component to automatically choose the optimal reading for each sentence. The f-structure output representation produced by the parser is converted into Prolog format before being passed to the transfer system, which is used both for augmenting the representation with lexicosemantic information and for applying the linguistic tests.

2.1. Lexical resources

2.1.1. Dorr's LCS database

Lexical conceptual structures (LCSs) from a database of 4269 English verbs (Dorr, 2001) are used to determine lexical aspect/verb type for each verbal predicate represented in the database (Dorr and Olsen, 1997).[2]

LCS representations encode the verbal semantics of their predicates in a directed-graph form that combines semantic structure and semantic content. The semantic structure is specified by the shape of the graph and its structural primitives and fields, and the content is specified through constants. Dorr and Olsen identify patterns within these representations which correspond to particular values for particular aspectual features. To extract this information, we performed pattern-matching searches over the entire database. Of 4269 predicates with 9806 readings, 219 verbs were identified as unambiguously stative, and 310 verbs as ambiguous between state and event. 1399 verbs were identified as unambiguously telic, and 1030 as ambiguous between telic and atelic.

2.1.2. Factive and propositional predicates

A second lexical resource, based on the discussion in (Peterson, 1997), identifies an additional two classes of predicates: factive and propositional. These predicates, when combined with clausal complements, result in the introduction of abstract entities to the discourse.

(3) factive: John knows that Mary won the race.

(4) propositional: John believes that all oxen should roam free.

Once extracted from the lexical resources, information about predicate-type is introduced into the f-structure (via transfer rules) as an additional feature attached to the predicate. For example, an instance of the predicate 'know' as a verb with a clausal complement would be augmented with the attribute-value combination 'pred-type(factive)'.

Two points should be noted with respect to augmentation of the f-structure with lexicosemantic information. First, the extraction and encoding of this information is a one-time event. Once encoded as transfer rules, the lexical information is invoked only when the relevant predicates appear in the packed Prolog term. Second, the repository of lexical information developed through this process is open-ended. Adding new resources or refining existing resources is a straightforward matter.

2.2. Term-rewriting transfer

Once the feature-structure for a sentence has been augmented with lexical information, the Situation Entity Evaluation Module (SEEM) applies a series of linguistic tests using PARC's term-rewriting transfer mechanism, originally developed for Machine Translation (Frank, 1999). When a situation entity is identified and classified, the transfer mechanism adds a situation entity (SE) feature to

[2] The database representations include WordNet senses and PropBank frames. We anticipate making use of this additional lexical information in future work.

```
"Lions sleep in the shade."

   PRED      'sleep<[30:lion]>'
             PRED   'lion'
   SUBJ      NTYPE  [GRAIN count]
           30CASE nom, NUM pl, PERS 3
                PRED   'in<[93:shade]>'
                       PRED   'shade'
                       NTYPE  [GRAIN unspecified]
   ADJUNCT  {   OBJ
                       SPEC   DET  [PRED 'the'
                                   [DET-FORM the, DET-TYPE def]
                     93CASE acc, NUM sg, PERS 3
              74ADV-TYPE vpadv, PSEM loc, PTYPE sem              }
   TNS-ASP  [MOOD indicative, PERF -_, PROG -_, TENSE pres]
 45PASSIVE -, STMT-TYPE decl, VTYPE main
```

Figure 1: F-structure representation for (2)

the f-structure for the appropriate predicate. The Prolog representation of the f-structure is then passed to a Discourse Mode Evaluation Module (DMEM) which matches the number and types of situation entities in the text passage to a set of criteria for the various discourse modes. Because there is no clear definition of the scope of this division, and no method has yet been determined for the automatic segmentation of texts into discourse substructures (though see (Passonneau and Litman, 1997)), for these experiments texts were segmented by paragraph, and discourse mode calculated for each paragraph in turn.

As mentioned above, the statistical disambiguation component of the Pargram grammar is used to select the most probable reading for each input sentence. This reading is then transformed into a packed Prolog term, representing the feature-structure as a flat set of descriptions in an approach inspired at least in part by the Shake ’n’ Bake method of machine translation (Whitelock, 1992). This flat representation allows us to identify linguistic features in their local context, without having to specify precise structural locations. Unless marked with a ’+’ preceding the feature in the rule’s left-hand side, features triggering the application of a rule are removed from the Prolog term once the rule has been applied. In this way, subsequent rules are prevented from applying to predicates which have already been analyzed.

To illustrate, we give the example of the identification of generic predicates via the linguistic correlate of a bare plural subject, as shown in (2) above. To implement this in the transfer system, we first use a rule which identifies and marks bare plural NP constructions.

```
+num(X,pl), +ntype(X,_), -spec(X,_)
==>
bplural(X,+).
```

Figure 2: Transfer rule for identification of bare plurals

The system scans the Prolog term for nominal predicators with a plural value for the number feature and nothing in the specifier position (the latter is indicated by the ’-’ marked feature in the left-hand side of the rule). When the transfer system finds a predicate whose f-structure meets the three requirements, the rule triggers and a new feature (’bplural’) is introduced to the predicate’s f-structure and given a ’+’ value. This feature is then picked up by a later rule which looks for a positive bplural value in the f-structure for predicates in subject position.

```
+subj(X,S), -xcomp(_,X), +bplural(S,+),
not_yet_marked(X,+)
==>
se_type(X,SE), type(SE,gen_stat).
```

Figure 3: Rule for identification of generic construction

When such a predicate is located, its f-structure receives a new feature which marks the situation entity and its type. Non ’+’-marked features are removed from the Prolog term (i.e. in the example above, the feature ’not-yet-marked’ is removed from the f-structure for the predicate represented by the variable ’X’).

The key to our characterization of discourse entities is ordering of the tests. To date we have derived seventeen separate linguistic tests and have ranked them according to their strength as correlates to particular situation entities. When in conflict, results from higher-ranking tests are preferred to results from lower-ranking tests. Ordering is enforced by the feature ’not-yet-marked’, which is added to the f-structures of all predicates in a pre-processing step. This feature is required in order for any rule to apply. Once a predicate has been assigned a situation entity type, the ’not-yet-marked’ feature is removed from that predicate’s f-structure, and the predicate is no longer available for analysis. Post-processing steps then remove unnecessary features from the Prolog term, which can be read back into the parser to produce an output representation in LFG f-structure form, as shown below.

3. Experiments and evaluation

The aim in this paper is to evaluate the effectiveness of the inclusion of lexical information in the output parse

```
"Translation of: Lions sleep in the shade."

    [PRED     'sleep<[-4-SUBJ:lion]>'                            ]
    [SUBJ     [PRED     'lion']                                  ]
    [         [BPLURAL +      ]                                  ]
    [                                                            ]
    [ADJUNCT { [PRED 'in<[-3-OBJ:shade]>']  }                    ]
    [        { -3[OBJ   [PRED 'shade']  ]  }                     ]
    [                                                            ]
  -4[SE-TYPE [TYPE gen_stat]                                     ]

  -2[PRED 'the']

  -1<GenGoodPunct>
```

Figure 4: Output representation for (2)

provided by the LFG-based grammar. To analyze this, we used two sets of transfer rules. The first set of rules employs only straightforward syntactic and lexical tests which apply to the unaugmented f-structures output by the parser. The bare plural test shown above is included in this basic set of transfer rules, as are rules identifying particular predicates associated with generics ('extinct', etc.) and rules using mood and aspectual features (e.g. progressive aspect, imperative mood) to identify eventualities. The second set is augmented with lexicosemantic information as described in 2.1.

We chose three selections from National Geographic magazine – two short (800-1000 word) articles, and the opening segment (of approximately the same length) of a longer article. No alterations other than preprocessing were made to the texts. We then compared the results from the parsing and analysis were compared to the human-annotated gold standard. According to the gold standard, Text1 contains 52 SEs, Text2 contains 85 SEs, and Text3 contains 59 SEs.

While inclusion of lexical information improved recall in the situation entity classification task, it decreased precision We find in general that the fuller set of transfer rules tends to overgenerate situation entities in some cases. It should also be noted that recall and precision are both improved by using the three-way distinction between situation entities, and that to get more useful input for the ultimate task of discourse mode calculation, the system will need to produce a more fine-grained analysis of situation entities. In the figure below, "nolex" refers to the basic set of transfer rules, and "withlex" to the expanded system.

Text	Rule Set	Recall	Precision
text1	nolex	46.2%	68.6%
	withlex	48.1%	61.0%
text2	nolex	70.6%	80.0%
	withlex	72.9%	77.5%
text3	nolex	52.5%	60.8%
	withlex	57.6%	57.6%
Avg.	nolex	56.4%	69.8%
	withlex	59.6%	65.4%

Figure 5: System evaluation

The results reported here are still considered to be preliminary. We expect to improve system performance by analyzing these results for refinement and expansion of the transfer rule set. Future work includes implementing a more fine-grained analysis of situation entities as well as incorporating additional sources of lexicosemantic information. We would also like to examine system performance with various combinations of resources, and to that end have designed the rule set in a modular fashion. Lexical resources can be combined in various configurations, which will allow us to analyze the impact of particular linguistic correlates and particular types of lexical information on the situation entity classification task.

4. Conclusions

In this paper we have reported on work in progress using multiple linguistic resources for the automatic classification of situation entities in naturally-occurring written English text. The basic architecture outlined in this paper will allow us to readily undertake research on the contributions of various linguistic phenomena and lexical information to the task of identifying and classifying situation entities.

5. Acknowledgements

The research for the project report was supported in part by a grant for linguistic research on text interpretation given to Dr. Carlota Smith by the New York Community Trust.

6. References

Carlson, G.N. and F.J. Pelletier (eds.), 1995. *The Generic Book*. Chicago: University of Chicago Press.

Dorr, Bonnie and Mari Olsen, 1997. Deriving verbal and compositional lexical aspect for NLP applications. In *Proceedings of ACL35*.

Dorr, Bonnie J., 2001. LCS verb database. University of Maryland. www.umiacs.umd.edu/˜bonnie/ LCS_Database_Documentation.html.

Frank, Anette, 1999. From parallel grammar development towards machine translation. a project overview. In *Proceedings of Machine Translation Summit VII*.

Passonneau, Rebecca and Diane Litman, 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139. Special Issue on Empirical Studies in Discourse Interpretation and Generation.

Peterson, Philip (ed.), 1997. *Fact Proposition Event*. Kluwer.

Riezler, Stefan, Dick Crouch, Ron Kaplan, Tracy King, John Maxwell, and Mark Johnson, 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Pennsylvania, Philadelphia*.

Smith, Carlota, 2003. *Modes of Discourse*. Cambridge University Press.

Whitelock, P., 1992. Shake-and-bake translation. In *Proceedings of COLING-92*. Nantes.