A Sequencing Model for Situation Entity Classification

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith
Department of Linguistics
University of Texas at Austin
{alexispalmer,ponvert,jbaldrid,carlotasmith}@mail.utexas.edu

Abstract

Situation entities (SEs) are the events, states, generic statements, and embedded facts and propositions introduced to a discourse by clauses of text. We report on the first data-driven models for labeling clauses according to the type of SE they introduce. SE classification is important for discourse mode identification and for tracking the temporal progression of a discourse. We show that (a) linguistically-motivated cooccurrence features and grammatical relation information from deep syntactic analysis improve classification accuracy and (b) using a sequencing model provides improvements over assigning labels based on the utterance alone. We report on genre effects which support the analysis of discourse modes having characteristic distributions and sequences of SEs.

1 Introduction

Understanding discourse requires identifying the participants in the discourse, the situations they participate in, and the various relationships between and among both participants and situations. Coreference resolution, for example, is concerned with understanding the relationships between references to discourse participants. This paper addresses the problem of identifying and classifying references to situations expressed in written English texts.

Situation entities (SEs) are the events, states, generic statements, and embedded facts and propositions which clauses introduce (Vendler, 1967; Verkuyl, 1972; Dowty, 1979; Smith, 1991; Asher, 1993; Carlson and Pelletier, 1995). Consider the text passage below, which introduces an *event*-type entity in (1), a *report*-type entity in (2), and a *state*-type entity in (3).

(1) Sony Corp. has heavily promoted the Video Walkman since the product's introduction last summer ,

(2) but Bob Gerson , video editor of This Week in Consumer Electronics , says

(3) Sony conceives of 8mm as a "family of products , camcorders and VCR decks , "

SE classification is a fundamental component in determining the discourse mode of texts (Smith, 2003) and, along with aspectual classification, for temporal interpretation (Moens and Steedman, 1988). It may be useful for discourse relation projection and discourse parsing.

Though situation entities are well-studied in linguistics, they have received very little computational treatment. This paper presents the first data-driven models for SE classification. Our two main strategies are (a) the use of linguistically-motivated features and (b) the implementation of SE classification as a sequencing task. Our results also provide empirical support for the very notion of discourse modes, as we see clear genre effects in SE classification.

We begin by discussing SEs in more detail. Section 3 describes our two annotated data sets and provides examples of each SE type. Section 4 discusses feature sets, and sections 5 and 6 present models, experiments, and results.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 896–903,
Prague, Czech Republic, June 2007.

2 Discourse modes and situation entities

In this section, we discuss some of the linguistic motivation for SE classification and the relation of SE classification to discourse mode identification.

2.1 Situation entities

The categorization of SEs into aspectual classes is motivated by patterns in their linguistic behavior. We adopt an expanded version of a paradigm relating SEs to discourse mode (Smith, 2003) and characterize SEs with four broad categories:

1. **Eventualities**. *Events* (E), particular *states* (S), and *reports* (R). R is a sub-type of E for SEs introduced by verbs of speech (e.g., *say*).

2. **General statives**. *Generics* (G) and *generalizing sentences* (GS). The former are utterances predicated of a general class or kind rather than of any specific individual. The latter are habitual utterances that refer to ongoing actions or properties predicated of specific individuals.

3. **Abstract entities**. *Facts* (F) and *propositions* (P).[1]

4. **Speech-act types**. *Questions* (Q) and *imperatives* (IMP).

Examples of each SE type are given in section 3.2.

There are a number of linguistic tests for identifying situation entities (Smith, 2003). The term *linguistic test* refers to a rule which correlates an SE type to particular linguistic forms. For example, event-type verbs in simple present tense are a linguistic correlate of GS-type SEs.

These linguistic tests vary in their precision and different tests may predict different SE types for the same clause. A rule-based implementation using them to classify SEs would require careful rule ordering or mediation of rule conflicts. However, since these rules are exactly the sort of information extracted as features in data-driven classifiers, they can be cleanly integrated by assigning them empirically determined weights. We use maximum entropy models (Berger et al., 1996), which are particularly well-suited for tasks (like ours) with many overlapping features, to harness these linguistic insights by using features in our models which encode, directly or indirectly, the linguistic correlates to SE types. The features are described in detail in section 4.

2.2 Basic and derived situation types

Situation entities each have a *basic situation type*, determined by the verb plus its arguments, the verb constellation. The verb itself plays a key role in determining basic situation type but it is not the only factor. Changes in the arguments or tense of the verb sometimes change the basic situation types:

(4) Mickey painted the house. (E)

(5) Mickey paints houses. (GS)

If SE type could be determined solely by the verb constellation, automatic classification of SEs would be a relatively straightforward task. However, other parts of the clause often override the basic situation type, resulting in aspectual coercion and a *derived situation type*. For example, a modal adverb can trigger aspectual coercion:

(6) Mickey probably paints houses. (P)

Serious challenges for SE classification arise from the aspectual ambiguity and flexibility of many predicates as well as from aspectual coercion.

2.3 Discourse modes

Much of the motivation of SE classification is toward the broader goal of identifying discourse modes, which provide a linguistic characterization of textual passages according to the situation entities introduced. They correspond to intuitions as to the rhetorical or semantic character of a text. Passages of written text can be classified into modes of discourse – Narrative, Description, Argument, Information, and Report – by examining concrete linguistic cues in the text (Smith, 2003). These cues are of two forms: the distribution of situation entity types and the mode of progression (either temporal or metaphorical) through the text.

[1] In our system these two SE types are identified largely as complements of factive and propositional verbs as discussed in Peterson (1997). Fact and propositional complements have some linguistic as well as some notional differences. Facts may have causal effects, and facts are in the world. Neither of these is true for propositions. In addition, the two have somewhat different semantic consequences of a presuppositional nature.

For example, the Narration and Report modes both contain mainly events and temporally bounded states; they differ in their principles of temporal progression. Report passages progress with respect to (deictic) speech time, whereas Narrative passages progress with respect to (anaphoric) reference time. Passages in the Description mode are predominantly stative, and Argument mode passages tend to be characterized by propositions and Information mode passages by facts and states.

3 Data

This section describes the data sets used in the experiments, the process for creating annotated training data, and preprocessing steps. Also, we give examples of the ten SE types.

There are no established data sets for SE classification, so we created annotated training data to test our models. We have annotated two data sets, one from the Brown corpus and one based on data from the Message Understanding Conference 6 (MUC6).

3.1 Segmentation

The Brown texts were segmented according to SE-containing clausal boundaries, and each clause was labeled with an SE label. Segmentation is itself a difficult task, and we made some simplifications. In general, clausal complements of verbs like *say* which have clausal direct objects were treated as separate clauses and given an SE label. Clausal complements of verbs which have an entity as a direct object and second clausal complement (such as *notify*) were not treated as separate clauses. In addition, some modifying and adjunct clauses were not assigned separate SE labels.

The MUC texts came to us segmented into elementary discourse units (EDUs), and each EDU was labeled by the annotators. The two data sets were segmented according to slightly different conventions, and we did not normalize the segmentation. The inconsistencies in segmentation introduce some error to the otherwise gold-standard segmentations.

3.2 Annotation

Each text was independently annotated by two experts and reviewed by a third. Each clause was assigned precisely one SE label from the set of ten possible labels. For clauses which introduce more than one SE, the annotators selected the most salient one. This situation arose primarily when complement clauses were not treated as distinct clauses, in which case the SE selected was the one introduced by the main verb. The label N was used for clauses which do not introduce any situation entity.

The Brown data set consists of 20 "popular lore" texts from section **cf** of the Brown corpus. Segmentation of these texts resulted in a total of 4390 clauses. Of these, 3604 were used for training and development, and 786 were held out as final testing data. The MUC data set consists of 50 Wall Street Journal newspaper articles segmented to a total of 1675 clauses. 137 MUC clauses were held out for testing. The Brown texts are longer than the MUC texts, with an average of 219.5 clauses per document as compared to MUC's average of 33.5 clauses. The average clause in the Brown data contains 12.6 words, slightly longer than the MUC texts' average of 10.9 words.

Table 1 provides examples of the ten SE types as well as showing how clauses were segmented. Each SE-containing example is a sequence of EDUs from the data sets used in this study.

SE	Text
S	That compares with roughly paperback-book dimensions for VHS.
G	Accordingly, most VHS camcorders are usually bulky and weigh around eight pounds or more.
S	"Carl is a tenacious fellow,"
R	said a source close to USAir.
GS	"He doesn't give up easily
GS	and one should never underestimate what he can or will do."
S	For Jenks knew
F	that Bari's defenses were made of paper.
E	Mr. Icahn then proposed
P	that USAir buy TWA,
IMP	"Fermate"!
R	Musmanno bellowed to his Italian crewmen.
Q	What's her name?
S	Quite seriously, the names mentioned as possibilities were three male apparatchiks from the Beltway's Democratic political machine
N	By Andrew B. Cohen Staff Reporter of The WSJ

Table 1: Example clauses and their SE annotation. Horizontal lines separate extracts from different texts.

W	
WORDS	words & punctuation
WT	
W	(see above)
POSONLY	POS tag for each word
WORD/POS	word/POS pair for each word
WTL	
WT	(see above)
FORCEPRED	T if clause (or preceding clause) contains *force* predicate
PROPPRED	T if clause (or preceding clause) contains propositional verb
FACTPRED	T if clause (or preceding clause) contains factive verb
GENPRED	T if clause contains generic predicate
HASFIN	T if clause contains finite verb
HASMODAL	T if clause contains modal verb
FREQADV	T if clause contains frequency adverb
MODALADV	T if clause contains modal adverb
VOLADV	T if clause contains volitional adverb
FIRSTVB	lexical item and POS tag for first verb
WTLG	
WTL	(see above)
VERBS	all verbs in clause
VERBTAGS	POS tags for all verbs
MAINVB	main verb of clause
SUBJ	subject of clause (lexical item)
SUPER	CCG supertag

Table 2: Feature sets for SE classification

3.3 Preprocessing

The linguistic tests for SE classification appeal to multiple levels of linguistic information; there are lexical, morphological, syntactic, categorial, and structural tests. In order to access categorial and structural information, we used the C&C[2] toolkit (Clark and Curran, 2004). It provides part-of-speech tags and Combinatory Categorial Grammar (CCG) (Steedman, 2000) categories for words and syntactic dependencies across words.

4 Features

One of our goals in undertaking this study was to explore the use of linguistically-motivated features and deep syntactic features in probabilistic models for SE classification. The nature of the task requires features characterizing the entire clause. Here, we describe our four feature sets, summarized in table 2. The feature sets are additive, extending very basic feature sets first with linguistically-motivated features and then with deep syntactic features.

[2] `svn.ask.it.usyd.edu.ap/trac/candc/wiki`

4.1 Basic feature sets: W and WT

The WORDS (W) feature set looks only at the words and punctuation in the clause. These features are obtained with no linguistic processing.

WORDS/TAGS (WT) incorporates part-of-speech (POS) tags for each word, number, and punctuation mark in the clause and the word/tag pairs for each element of the clause. POS tags provide valuable information about syntactic category as well as certain kinds of shallow semantic information (such as verb tense). The tags are useful for identifying verbs, nouns, and adverbs, and the words themselves represent lexico-semantic information in the feature sets.

4.2 Linguistically-motivated feature set: WTL

The WORDS/TAGS/LINGUISTIC CORRELATES (WTL) feature set introduces linguistically-motivated features gleaned from the literature on SEs; each feature encodes a linguistic cue that may correlate to one or more SE types. These features are not directly annotated; instead they are extracted by comparing words and their tags for the current and immediately preceding clauses to lists containing appropriate triggers. The lists are compiled from the literature on SEs.

For example, clauses embedded under predicates like *force* generally introduce E-type SEs:

(7) I forced [John to run the race with me].

(8) * I forced [John to know French].

The feature `force-PREV` is extracted if a member of the *force*-type predicate word list occurs in the previous clause.

Some of the correlations discussed in the literature rely on a level of syntactic analysis not available in the WTL feature set. For example, stativity of the main verb is one feature used to distinguish between event and state SEs, and particular verbs and verb tenses have tendencies with respect to stativity. To approximate the main verb without syntactic analysis, WTL uses the lexical item of the first verb in the clause and the POS tags of all verbs in the clause.

These linguistic tests are non-absolute, making them inappropriate for a rule-based model. Our models handle the defeasibility of these correlations probabilistically, as is standard for machine learning for natural language processing.

4.3 Addition of deep features: WTLG

The WORDS/TAGS/LINGUISTIC CORRELATES/GRAMMATICAL RELATIONS (WTLG) feature set uses a deeper level of syntactic analysis via features extracted from CCG parse representations for each clause. This feature set requires an additional step of linguistic processing but provides a basis for more accurate classification.

WTL approximated the main verb by sloppily taking the first verb in the clause; in contrast, WTLG uses the main verb identified by the parser. The parser also reliably identifies the subject, which is used as a feature.

Supertags –CCG categories assigned to words– provide an interesting class of features in WTLG. They succinctly encode richer grammatical information than simple POS tags, especially subcategorization and argument types. For example, the tag S$\backslash$NP denotes an intransitive verb, whereas (S$\backslash$NP)/NP denotes a transitive verb. As such, they can be seen as a way of encoding the verbal constellation and its effect on aspectual classification.

5 Models

We consider two types of models for the automatic classification of situation entities. The first, a labeling model, utilizes a maximum entropy model to predict SE labels based on clause-level linguistic features as discussed above. This model ignores the discourse patterns that link multiple utterances. Because these patterns recur, a sequencing model may be better suited to the SE classification task. Our second model thus extends the first by incorporating the previous n ($0 \leq n \leq 6$) labels as features.

Sequencing is standardly used for tasks like part-of-speech tagging, which generally assume smaller units to be both tagged and considered as context for tagging. We are tagging at the clause level rather than at the word level, but the structure of the problem is essentially the same. We thus adapted the OpenNLP maximum entropy part-of-speech tagger[3] (Hockenmaier et al., 2004) to extract features from utterances and to tag sequences of utterances instead of words. This allows the use of features of adjacent clauses as well as previously-predicted labels when making classification decisions.

[3] `http://opennlp.sourceforge.net.`

6 Experiments

In this section we give results for testing on Brown data. All results are reported in terms of accuracy, defined as the percentage of correctly-labeled clauses. Standard 10-fold cross-validation on the training data was used to develop models and feature sets. The optimized models were then tested on the held-out Brown and MUC data.

The baseline was determined by assigning *S* (state), the most frequent label in both training sets, to each clause. Baseline accuracy was 38.5% and 36.2% for Brown and MUC, respectively.

In general, accuracy figures for MUC are much higher than for Brown. This is likely due to the fact that the MUC texts are more consistent: they are all newswire texts of a fairly consistent tone and genre. The Brown texts, in contrast, are from the 'popular lore' section of the corpus and span a wide range of topics and text types. Nonetheless, the patterns between the feature sets and use of sequence prediction hold across both data sets; here, we focus our discussion on the results for the Brown data.

6.1 Labeling results

The results for the labeling model appear in the two columns labeled 'n=0' in table 3. On Brown, the simple W feature set beats the baseline by 6.9% with an accuracy of 45.4%. Adding POS information (WT) boosts accuracy 4.5% to 49.9%. We did not see the expected increase in performance from the linguistically motivated WTL features, but rather a slight decrease in accuracy to 48.9%. These features may require a greater amount of training material to be effective. Addition of deep linguistic information with WTLG improved performance to 50.6%, a gain of 5.2% over words alone.

6.2 Oracle results

To determine the potential effectiveness of sequence prediction, we performed oracle experiments on Brown by including previous gold-standard labels as features. Figure 1 illustrates the results from oracle experiments incorporating from zero to six previous gold-standard SE labels (the *lookback*). The increase in performance illustrates the importance of context in the identification of SEs and motivates the use of sequence prediction.

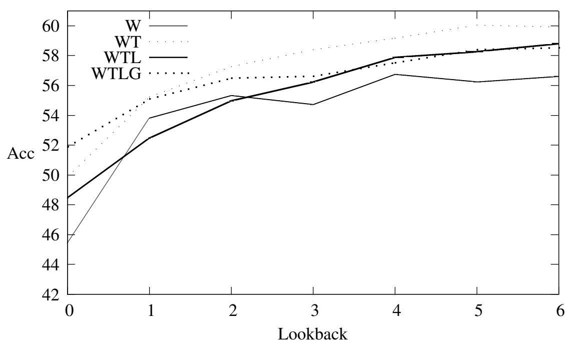

Figure 1: **Oracle results** on Brown data.

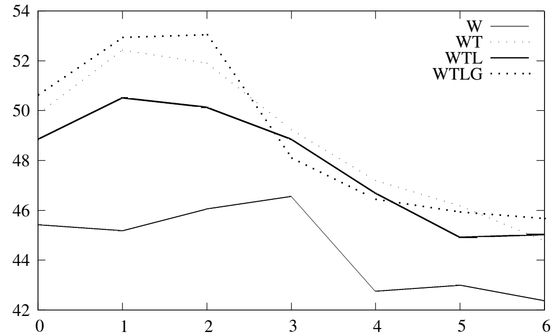

Figure 2: Sequencing results on Brown data.

6.3 Sequencing results

Table 3 gives the results of classification with the sequencing model on the Brown data. As with the labeling model, accuracy is boosted by WT and WTLG feature sets. We see an unexpected degradation in performance in the transition from WT to WTL.

The most interesting results here, though, are the gains in accuracy from use of previously-predicted labels as features for classification. When labeling performance is relatively poor, as with feature set W, previous labels help very little, but as labeling accuracy increases, previous labels begin to effect noticeable increases in accuracy. For the best two feature sets, considering the previous two labels raises the accuracy 2.0% and 2.5%, respectively.

In most cases, though, performance starts to degrade as the model incorporates more than two previous labels. This degradation is illustrated in Figure 2. The explanation for this is that the model is still very weak, with an accuracy of less than 54% for the Brown data. The more previous predicted labels the model conditions on, the greater the likelihood that one or more of the labels is incorrect. With gold-standard labels, we see a steady increase in accuracy as we look further back, and we would need a better performing model to fully take advantage of knowledge of SE patterns in discourse.

The sequencing model plays a crucial role, particularly with such a small amount of training material, and our results indicate the importance of local context in discourse analysis.

BROWN	Lookback (n)						
	0	1	2	3	4	5	6
W	45.4	45.2	46.1	46.6	42.8	43.0	42.4
WT	49.9	52.4	51.9	49.2	47.2	46.2	44.8
WTL	48.9	50.5	50.1	48.9	46.7	44.9	45.0
WTLG	50.6	52.9	**53.1**	48.1	46.4	45.9	45.7
Baseline	38.5						

Table 3: SE classification results with sequencing on **Brown test set**. Bold cell indicates accuracy attained by model parameters that performed best on development data.

6.4 Error analysis

Given that a single one of the ten possible labels occurs for more than 35% of clauses in both data sets, it is useful to look at the distribution of errors over the labels. Table 4 is a confusion matrix for the held-out Brown data using the best feature set.[4] The first column gives the label and number of occurrences of that label, and the second column is the accuracy achieved for that label. The next two columns show the percentage of erroneous labels taken by the labels S and GS. These two labels are the most common labels in the development set (38.5% and 32.5%). The final column sums the percentages of errors assigned to the remaining seven labels. As one would expect, the model learns the predominance of these two labels. There are a few interesting points to make about this data.

First, 66% of G-type clauses are mistakenly assigned the label GS. This is interesting because these two SE-types constitute the broader SE cat-

[4] Thanks to the anonymous reviewer who suggested this useful way of looking at the data.

Label	% Correct Label	% Incorrect S	GS	Other
S(278)	72.7	n/a	14.0	13.3
E(203)	50.7	37.0	11.8	0.5
GS(203)	44.8	46.3	n/a	8.9
R(26)	38.5	30.8	11.5	19.2
N(47)	23.4	31.9	23.4	21.3
G(12)	0.0	25.0	66.7	8.3
IMP(8)	0.0	75.0	25.0	0.0
P(7)	0.0	71.4	28.6	0.0
F(2)	0.0	100.0	0.0	0.0

Table 4: Confusion matrix for Brown held-out test data, WTLG feature set, lookback n = 2. Numbers in parentheses indicate how many clauses have the associated gold standard label.

WTLG	**lookback**	Brown test set
train:Brown	0	50.6
	2	53.1
train:MUC	0	43.1
	2	42.9
train:all	0	50.4
	2	49.5

Table 5: Cross-domain SE classification

egory of generalizing statives. The distribution of errors for R-type clauses points out another interesting classification difficulty.[5] Unlike the other categories, the percentage of *false-other* labels for R-type clauses is higher than that of *false-GS* labels. 80% of these *false-other* labels are of type E. The explanation for this is that R-type clauses are a subtype of the event class.

6.5 Genre effects in classification

Different text domains frequently have different characteristic properties. Discourse modes are one way of analyzing these differences. It is thus interesting to compare SE classification when training and testing material come from different domains.

Table 5 shows the performance on Brown when training on Brown and/or MUC using the WTLG feature set with simple labeling and with sequence prediction with a lookback of two. A number of things are suggested by these figures. First, the labeling model (lookback of zero), beats the baseline even when training on out-of-domain texts (43.1% vs. 38.5%), but this is unsurprisingly far below training on in-domain texts (43.1% vs. 50.6%). Second, while sequence prediction helps with in-domain training (53.1% vs 50.6%), it makes no difference with out-of-domain training (42.9% vs 43.1%). This indicates that the patterns of SEs in a text do indeed correlate with domains and their discourse modes, in line with case-studies in the discourse modes theory (Smith, 2003). Finally, mixing out-of-domain training material with in-domain material does not hurt labelling accuracy (50.4% vs 50.6%), but it does take away the gains from sequencing (49.5% vs 53.1%).

These genre effects are suggestive, but inconclusive. A similar setup with much larger training and testing sets would be necessary to provide a clearer picture of the effect of mixed domain training.

7 Related work

Though we are aware of no previous work in SE classification, others have focused on automatic detection of aspectual and temporal data.

Klavans and Chodorow (1992) laid the foundation for probabilistic verb classification with their interpretation of aspectual properties as gradient and their use of statistics to model the gradience. They implement a single linguistic test for stativity, treating lexical properties of verbs as tendencies rather than absolute characteristics.

Linguistic indicators for aspectual classification are also used by Siegel (1999), who evaluates 14 indicators to test verbs for stativity and telicity. Many of his indicators overlap with our features.

Siegel and McKeown (2001) address classification of verbs for stativity (event vs. state) and for completedness (culminated vs. non-culminated events). They compare three supervised and one unsupervised machine learning systems. The systems obtain relatively high accuracy figures, but they are domain-specific, require extensive human supervision, and do not address aspectual coercion.

Merlo and Stevenson (2001) use corpus-based thematic role information to identify and classify unergative, unaccusative, and object-drop verbs. Stevenson and Merlo note that statistical analysis cannot and should not be separated from deeper linguistic analysis, and our results support that claim.

[5]Thanks to an anonymous reviewer for bringing this to our attention.

The advantages of our approach are the broadened conception of the classification task and the use of sequence prediction to capture a wider context.

8 Conclusions

Situation entity classification is a little-studied but important classification task for the analysis of discourse. We have presented the first data-driven models for SE classification, motivating the treatment of SE classification as a sequencing task.

We have shown that linguistic correlations to situation entity type are useful features for probabilistic models, as are grammatical relations and CCG supertags derived from syntactic analysis of clauses. Models for the task perform poorly given very basic feature sets, but minimal linguistic processing in the form of part-of-speech tagging improves performance even on small data sets used for this study. Performance improves even more when we move beyond simple feature sets and incorporate linguistically-motivated features and grammatical relations from deep syntactic analysis. Finally, using sequence prediction by adapting a POS-tagger further improves results.

The tagger we adapted uses beam search; this allows tractable use of maximum entropy for each labeling decision but forgoes the ability to find the *optimal* label sequence using dynamic programming techniques. In contrast, Conditional Random Fields (CRFs) (Lafferty et al., 2001) allow the use of maximum entropy to set feature weights with efficient recovery of the optimal sequence. Though CRFs are more computationally intensive, the small set of SE labels should make the task tractable for CRFs.

In future, we intend to test the utility of SEs in discourse parsing, discourse mode identification, and discourse relation projection.

Acknowledgments

This work was supported by the Morris Memorial Trust Grant from the New York Community Trust. The authors would like to thank Nicholas Asher, Pascal Denis, Katrin Erk, Garrett Heifrin, Julie Hunter, Jonas Kuhn, Ray Mooney, Brian Reese, and the anonymous reviewers.

References

N. Asher. 1993. *Reference to Abstract objects in Discourse*. Kluwer Academic Publishers.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

G. Carlson and F. J. Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press, Chicago.

S. Clark and J. R. Curran. 2004. Parsing the WSJ using CCG and log–linear models. In *Proceedings of ACL–04*, pages 104–111, Barcelona, Spain.

D. Dowty. 1979. *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

J. Hockenmaier, G. Bierner, and J. Baldridge. 2004. Extending the coverage of a CCG system. *Research on Language and Computation*, 2:165–208.

J. L. Klavans and M. S. Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of COLING 14*, Nantes, France.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of ICML*, pages 282–289, Williamstown, USA.

P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*.

M. Moens and M. Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

P. Peterson. 1997. *Fact Proposition Event*. Kluwer.

E. V. Siegel and K. R. McKeown. 2001. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628.

E. V. Siegel. 1999. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of ACL37*, University of Maryland, College Park.

C. S. Smith. 1991. *The Parameter of Aspect*. Kluwer.

C. S. Smith. 2003. *Modes of Discourse*. Cambridge University Press.

M. Steedman. 2000. *The Syntactic Process*. MIT Press/Bradford Books.

Z. Vendler, 1967. *Linguistics in Philosophy*, chapter Verbs and Times, pages 97–121. Cornell University Press, Ithaca, New York.

H. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. Reidel, Dordrecht.