Identification of Caused Motion Constructions

Jena D. Hwang
University of Colorado at Boulder
Boulder, CO 80309
hwangd@colorado.edu

Martha Palmer
University of Colorado at Boulder
Boulder, CO 80309
martha.palmer@colorado.edu

Abstract

This research describes the development of a supervised classifier of English Caused Motion Constructions (CMCs) (e.g. *The goalie kicked the ball into the field*). Consistent identification of CMCs is a necessary step to a correct interpretation of semantics for sentences where the verb does not conform to the expected semantics of the verb (e.g. *The crowd laughed the clown off the stage*). We expand on a previous study on the classification CMCs (Hwang et al., 2010) to show that CMCs can be successfully identified in the corpus data. In this paper, we present the classifier and the series of experiments carried out to improve its performance.

1 Introduction

While natural language processing performance has been improved through the recognition that there is a relationship between the semantics of the verb and the syntactic context in which the verb is realized (Guildea and Palmer, 2002), sentences where the verb does not conform to the expected syntax-semantic patterning behavior remain problematic.

1. The goalie kicked the ball into the field.
2. The crowd laughed the clown off the stage.

These sentences are semantically related – an entity causes a second entity to go along the path described by the prepositional phrase: in 1, the goalie causes the ball to go into the field, and in 2, the crowd causes the clown to go off the stage.

While only the verb in the first sentence is generally identified as a verb of motion that can appear in a caused motion context, both are examples of caused motion constructions (CMCs) (Goldberg, 1995). The verb *laugh* of sentence 2 is normally considered an intransitive manner of speaking verb (e.g. *The crowd laughed at the clown*), but in this sentence, the verb is coerced into the caused motion interpretation and the semantics of the verb gives the manner in which the movement happened (e.g. *the crowd caused the clown to move off the stage by means of laughing*). The semantics parallel one another: both sentences have a causal argument responsible for the event, an argument in motion, and a path that specifies the initial, middle, or final location, state or condition of the argument in motion (Hwang et al., 2013).

Thus, if the semantic interpretation is strictly based on the expected semantics of the verb and its arguments, it fails to include the relevant information from the CMC. Accurate semantic role labelling requires that NLP classifiers accurately identify these coerced usages in data.

In a previous study, we carried out preliminary work on the supervised identification of CMCs (Hwang et al., 2010). The pilot study was conducted in a highly controlled environment over a small portion of Wall Street Journal (WSJ) data. The annotation of CMCs were limited to 1.8K instances of WSJ data. In the pilot, we were able to establish a classifier predicting CMC with high accuracy (87.2% precision, 86.0% recall, and 0.866 f-score).

In a subsequent study, we developed a detailed set of criteria for identifying CMCs to insure the

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 51–60,
Denver, Colorado, June 4–5, 2015.

production of consistent annotation with high inter-annotator agreement (Hwang et al., 2014). Through the semantic typing of the CMCs, the annotation guidelines defining CMCs were further refined from the guidelines used during the pilot study. Using the newly established criteria for annotation, we extended the annotation over the complete WSJ, and further included the Broadcast News and Webtext for the annotation of CMC. This study resulted in over 20K instances of CMC annotation.

In this paper, we carry out a supervised classification of the CMC. This study further expands on a pilot study with the larger set of high-quality annotated data for the further training and testing of CMC classifiers.

2 Caused Motion Constructions

CMCs are defined as having the coarse-grained syntactic structure of Subject Noun Phrase followed by a verb that takes both a Noun Phrase Object and a Prepositional Phrase: (NP-SBJ (V NP PP)); and the semantic meaning 'The agent, NP-SBJ, directly causes the patient, NP, to move along the path specified by the PP' (Goldberg, 1995). This construction is exemplified by the following sentences:

3. Frank sneezed the tissue off the table.
4. John stuffed the letter in the envelope.
5. Sally threw a ball to him.

However, not all syntactic structures of the form (NP-SBJ (V NP PP)):

6. Mary kicked the ball to my relief.
7. Jen took the highway into Pennsylvania.
8. We saw the bird in the shopping mall.

In 6, the PP does not specify a direction or a path. In 8, PP indicates the location in which the "seeing" event happened, not a path along which "we" caused "the bird" to move. Though the PP in 7 expresses a path, it is not a path over which Jen causes "the highway" to move.

3 Experimental Setup

3.1 Corpora

Our data comes from the latest version of OntoNotes, version 5.0, (Weischedel et al., 2012). Gold annotations for Penn Treebank, PropBank, and Verb Sense Annotation are available for all of OntoNotes corpora. As we did for the pilot study, we use the Wall Street Journal (WSJ) corpus. This corpus contains over 846K words selected from the non "strictly" financial (e.g., daily market reports) portion of the Wall Street Journal included in the Penn Treebank II (Marcus et al., 1994). We also pull from the smaller of the two WebText (WEB) data sets published in OntoNotes. This corpus contains 85K words selected from English weblogs. This portion of the data is not to be confused with the the larger 200K word web data, which is a separate corpus in OntoNotes. The third corpus used in our experiments is the 200K word Broadcast News (BN) data. OntoNotes' BN data contains news texts from broadcasting sources such as CNN, ABC, and PRI (Public Radio International).

3.2 Data Selection

In order to narrow the data down to a more manageable size for annotation, we exclude instances that can be deterministically categorized as NON-CMCs using the gold Penn Treebank annotation of the corpora. To do this we first select all sentences with the base syntactic form (NP-SBJ (V NP PP)) based on the Penn Treebank gold annotation.

Additionally, we use a set of heuristics (a smaller set than the pilot) to further select instances of potential CMCs. Instances which satisfy the following three conditions are extracted for annotation:(1) an NP exists in the verb phrase; (2) at least one PP exists in the verb phrase; and (3) the NP precedes the PP in the verb phrase.

For the remaining data, already annotated instances from the pilot study are separated out for double-checking. We also set aside instances that can be deterministically categorized as NON-CMC: instances with the function tags ADV, EXT, PRD, VOC, or TMP. These sentences are kept for a quick verification at the annotation stage that they indeed are cases of NON-CMCs and labeled as such.

3.3 Added Syntactic Complexity

In the pilot study, we had excluded passive instances (e.g. *Coffee was shipped from Colombia by Gracie.*), instances with traces in the object NP or PP including questions, relative clauses, and subordinate clauses (e.g. *What did Gracie ship from Colombia?* and *It was Gracie that shipped coffee from Colombia.*) and instances in which the verb is a conjunct to the main verb in the sentence (e.g. *chop* in *He peeled the potatoes and chopped them into a bowl*), opting to match sentences by their surface structure. For the current study, our data selection includes instances that retain an underlying syntactic form (NP-SBJ (V NP PP)). In effect, we extend the syntactic variability in the data.

Form	WSJ	BN	WEB
Questions/ Rel. clauses	2.3%	3.9%	2.6%
Passives	4.4%	4.6%	1.6%
Conjuncts	7.9%	10.2%	16.3%
Other clauses	46.3%	41.2%	37.3%
Other	41.4%	44.1%	44.7%

Table 1: Syntactic forms found in data. Other clauses include both subordinate and complement clauses.

Table 1 shows the breakdown of the syntactic forms in the current data. The pilot data was solely restricted to the "Other" category. More than half of all the syntactic forms represented in our current data add to the syntactic complexity beyond that of the pilot dataset, and lower our baseline classifier performance significantly.

3.4 Labels and Classfiers

The annotated data includes 4 major types of CMCs (Hwang et al., 2014). CMC types are listed below:

- **Displacement:** These CMCs express a (concrete or abstract) change of location of an entity (e.g. *The goalie kicked the ball into the field.* or *The market tilted the economy into recession.*). This is the most prototypical CMC type.
- **Change of Scale:** These CMCs express a change in value on a linear scale (e.g. *Torrential rains raised the water level to 500ft.*).
- **Change of Possesion:** These CMCs express a change of possession (e.g. *John gave a book to Mary*).
- **Change of State:** These CMCs express a change of attribute of an item (e.g. *I smashed the vase into pieces.*)

The experiments presented in this paper are geared towards the identification of: (1) all 4 types unified under a single label and (2) the "Displacement" type of CMCs (1 of the 4 types). We build two binary classifiers – one for each of the two labels. We will refer to the former classifier as "CMC classifier" and the latter as the "DISPLACE classifier". Table 2 shows the classification label distribution across the three corpora.

For all our experiments, 80% of the annotated data is randomly selected as the training/development data and the remaining 20% is set aside as the test/evaluation set. For our experiments, we use a Support Vector Machine (SVM) classifier with a linear kernel. In particular, we use LIBSVM (Chang and Lin, 2001) as our training and testing software. We use a 5-fold cross-validation process for the development stage.

3.5 Features

The features encode syntactic and semantic information that targets four elements in the sentence: (1) the verb, which expresses the event or the situation of the sentence, (2) the preposition, which instantiates the path information in a caused motion sentence, (3) the complement of the preposition, which covers the rest of the prepositional phrase, (4) the cause argument, which is recovered from the subject of the sentence or the prepositional by-phrase in a passive sentence, and (5) the undergoer argument, which is recovered from the direct object position of the sentence or from the subject position in a passive sentence. We will discuss the cause and undergoer argument recovery in further detail later.

3.5.1 Feature Sets

The **baseline** feature set is encoded by the **verb lemma** – the lemmatized and case-normalized verb. The verb lemma feature is the baseline feature for all our experiments. Following are the semantic and syntactic features sets used in our experiments. Anytime we use the terms "Full Set" or full feature set, we are referring to a set of features that includes all of the feature sets below for each of the four

	WSJ		WEB		BN	
CMC	2250	14.8%	533	29.2%	703	18.6%
NONCMC	12959	85.2%	1291	70.8%	3073	81.4%
DISPLACE	1261	8.3%	412	22.6%	511	13.5%
NONDISPLACE	13948	91.7%	1412	77.4%	3265	86.5%

Table 2: CMC and DISPLACE label distribution in training and test data

elements as noted above.

Features encoding **semantic** information are as following:

- **Nominal Entity** features which are automatically generated using BBNs IdentiFinder (Bikel et al., 1999). The IdentiFinder annotates relevant noun phrases with labels such as "Persons", "Time", "Location", or "Organization".
- **PropBank Frameset** features specify the verb's sense based on its subcategorization frame. This is extracted from the gold annotation provided by Ontonotes.
- **Ontonotes Verb Sense** features which specify the verb's sense. The semantics of these features are generally finer grained than what the PropBank framesets encode. These features are also provided as gold annotation in OntoNotes.
- **VerbNet Class** features that encode each of the VerbNet classes in which the verb is a member. A verb can be a member of one or more classes.
- **Preposition Type** features obtained from the automatic preposition labeller developed in a recent study by (Srikumar, 2013). The labeller introduces a set of 32 roles to disambiguate semantics of prepositions as used in sentences (e.g. *from* in *Her sudden death* ***from*** *pneumonia ...* (Cause) vs. *She copied the lines* ***from*** *the film.*(Source))

Features encoding **syntactic** information include:

- **Part of Speech Tag** of the lexical item in the syntactic parse.
- **Dependency Relation Tag** of the lexical item in a dependency parse.

Please note that while we depend on the phrasal trees for the data selection process, for feature extraction, we employ the CLEAR dependency parses (Choi, 2012). These parses have been automatically converted from the Penn Treebank phrasal trees. The decision to encode syntactic features from the dependency parses rather than from phrasal parses was based on the flexibility and the amount of additional information we gain through the dependency parse type. After a series of experimental runs with features from both parse types, it was determined that further syntactic features based on the phrase trees produced relatively similar performance to that of its counterpart labels on the dependency trees. However, the dependency labels are functionally finer grained than phrase structure labels for those syntactic elements that are most relevant to the CMCs.

3.5.2 Cause & Undergoer Argument Recovery

We make a pre-processing pass of the data to recover these arguments when possible. The recovered arguments are as following:

- **Passive Sentences:** For passive sentences, the complement of the *by*-prepositional phrase is recovered as the cause argument and the subject is recovered as the undergoer argument.
- **Conjunctions:** Given two verbal conjuncts sharing the subject, as in *"He cut the peppers and diced the tomatoes"*, the CLEAR dependency parse places the conjunction and the second conjunct as dependents of the first verb. This means that in dependency trees the two conjuncts' access to the cause argument is not symmetrical. The argument *He* is accessible to the verb *diced* via the verb *cut*, as the argument is a direct dependent of the verb *cut* and not the verb *diced*. To recover the arguments of the

	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Baseline	61.23	37.56	0.4656	75.6	55.7	0.641	71.4	53.6	0.612
Baseline+P	**75.00**	74.67	0.7483†	78.0	**80.2**	**0.791**†	**84.8**	75.7	0.800†
Full Set	74.00	**77.78**	**0.7584**†	**79.0**	78.3	0.787†	84.1	**82.9**	**0.835**†
Annotator Agreement			0.667			0.764			0.606

Table 3: System performance on CMC label classification.
Statistically significant change from the Baseline feature set is marked with a †.

	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Baseline	66.80	63.89	0.6531	72.7	58.5	0.649	71.3	55.9	0.626
Baseline+P	**76.33**	74.21	**0.7525**†	73.4	70.7	0.720	80.0	70.6	0.750†
Full Set	72.52	**75.40**	0.7393†	**76.5**	**79.3**	**0.778**†	**80.6**	**77.5**	**0.790**†

Table 4: System performance on DISPLACE label classification.
Statistically significant change from the Baseline feature set is marked with a †.

second verb conjunct we reach for the dependent on the first conjunct as necessary.

- **Subordinate clauses:** For verbs that are found in subordinate clauses whose head node is a verb (also called matrix verb) such as an infinitival clause (e.g. *He [plans]*-HEAD *to* ***cut*** *the peppers into pieces*), or a relative clause (e.g. *Joe [cut]*-HEAD *the tomatoes Mary* ***washed***.), we reach for the head node's arguments to fill in the missing cause and theme arguments. If there is an intervening relative pronoun (e.g. *Joe cut the tomatoes* ***that*** *Mary washed*), the relative pronoun is retrieved as the argument (either as cause or theme depending whether or not the subordinate clause is a passive), instead.

3.5.3 POS Tags & Dependency Relation Tags

After a series of experiments, it was determined that the part of speech and the dependency relation features might be too fine grained to provide useful information to the classifier. Thus, all of the features expressed by the part of speech and the dependency relation are featurized in the following manner.

- **Part of Speech Tags:** (1) Cardinal numbers (CD), pronouns (PRP), and gerundial (VBG) and participial (VBN) forms of verbs are featurized as found (one feature per tag). (2) Rest of the verb forms are mapped to the base tag VB. (3) Plural nouns are mapped to their singular counterparts. (4) Adjectives and adverbs are mapped to the base tag JJ and RB, respectively. (5) Rest are given the tag: OTHER.

- **Dependency Relation Labels:** (1) Relations specifying subjects, direct object, and agent (oblique of a passive sentence), and relations specifying the object of the preposition, complement clauses, and relative clauses are featurized as found (one feature per tag). (2) Complement clauses (e.g. *pcomp*, *acomp*) are grouped under a single *comp* label. (3) Modifiers (e.g. *partmod*, *advmod*) are grouped under the *mod* label. (4) Rest are given the tag: OTHER.

4 Classifier Experiments

Tables 3 and 4 show the precision and recall percentages and the f-score values for our experiments. Here we show results for three feature combinations: the **Baseline** set encoded from the verb's lemma, the **Baseline** plus the preposition feature set (**Baseline+P**), and the **Full Set** that includes all of the features listed in Section 3.5. The best performance values are bold-faced. The significance of a feature set's performance was evaluated via a chi-squared test (McNemar, $p < 0.05$). Statistically significant change from the **Baseline** feature set is marked with a †. Additionally, for the CMC classification we show the inter-annotator agreement

(Gold) f-score (Hwang et al., 2014). Our best performances in CMC classification as measured by the f-score are comparable or higher than the inter annotator agreement f-score.

4.1 Syntactic vs. Semantic Features

With the exception of the DISPLACE classifier on the WEB corpus, both the **Baseline+P** and the **Full Set** of features perform significantly better than the **Baseline** in both sets of experiments. It is interesting that the **Baseline+P** set performs just as well and sometimes better than the full set of feature consistently across the corpora, though the differences in the values are not statistically significant.

In order to gain a better understanding of the performance on the full set of features, the full feature set was divided into syntactic features and semantic features as described in Section 3.5. As a means of control, both the syntactic and semantic feature sets also include the features for the verb lemma and the preposition. Out of the different feature combinations examined, the distinction between semantic and syntactic features is the most salient. Table 5 shows the system performance values for the syntactic and semantic features. We also show the performance of the **Baseline+P** plus VerbNet class (**Baseline+PV**) feature set, as it gives better insight into the semantic feature performance.

The numbers indicate that the semantic features have a consistently higher performance than the syntactic features. The syntactic feature sets, perform significantly lower than the full feature sets and they barely pass the **Baseline** features in performance. In fact, the syntactic features are significantly lower than the **Baseline+P** features, despite the fact that, just like the semantic features, they include the verb lemma feature and the preposition feature. This suggests, that the syntactic features even in the presence of the lexical features are not strongly predictive of caused motion constructions. Moreover, these numbers seem to indicate that the performance on the full set of features likely comes from the semantic feature performance.

Amongst the semantic features, the **Baseline** feature, the **Baseline+P** feature, and the feature for VerbNet class membership of the verb (i.e. **Baseline+PV**) give the highest results. With the exception of the CMC classifier on the BN corpus, the numbers for the **Baseline+PV** set are not significantly different from either the semantic feature or the full feature set performance. Other semantic combinations were also tested, but they did not result in any particular change from the semantic feature set and the full feature set.

The semantic features perform as the most predictive features. This finding makes intuitive sense. Recall that during the data selection stage, we selected for instances that show syntactic compatibility with CMCs. Although syntactic variability still exists in the selected data (e.g. relative clauses and passive sentences), because of the data selection stage based on syntax, the task of identification comes primarily down to the semantic distinction between existing sentences. Additionally, some of the existing syntactic differences are neutralized by the cause and undergoer argument pre-processing stage described in Section 3.5.2. Thus, it stands to reason that most of the useful contributions come from the lexical items themselves and the semantics of the verb and its arguments.

Finally, the baseline system of the DISPLACE classification shows either a similar or improved performance over the CMC classifier. The overall performances across the different feature sets show similar values. Given that DISPLACE makes up a smaller percentage of the total data as shown in Section 3.4 (e.g. DISPLACE label for WSJ accounts for just under 9% of the total test and training data), the comparable performance is likely indicative that the DISPLACE label represents a more semantically coherent phenomenon than the CMC label.

4.2 Removing Frequent NON-CMC Verbs

In this experiment, we remove the top 25 highly frequent verbs[1] that do not appear in a CMC usage from both the training and testing data[2]. Their semantics are not compatible with the established definitions of CMCs. For example, verbs like *be*, *do*, or *have* cannot have caused motion usages, and verbs

[1] We effectively went down the list of the most frequent verbs in our WSJ data, and stopped at the first verb that could be judged as compatible and non-contrary to the established definitions of CMCs. 25 is the number of verbs in this list before the first CMC-compatible verb was reached.

[2] Top 25 verbs include: *accuse, base, be, build charge, create, do, fall, file, find, have, hold, keep, leave, offer, open, play, prevent, produce, quote, reach, rise, see, use*, and *view*.

CMC Classification									
	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Syntactic	63.79	41.11	0.5000	76.6	55.7	0.645	72.4	54.3	0.620
Semantic	71.02	72.44	0.7173	77.3	64.2	0.701	80.5	76.4	0.784
Baseline+PV	71.78	76.89	0.7425	78.8	77.4	0.781	85.9	82.9	0.844

DISPLACE Classification									
	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Syntactic	66.80	63.89	0.6531	73.8	58.5	0.653	72.3	58.8	0.649
Semantic	72.94	73.81	0.7337	76.3	70.7	0.734	74.3	79.4	0.768
Baseline+PV	74.81	76.59	0.7569	78.7	72.0	0.752	82.8	75.5	0.790

Table 5: System performance on semantic and syntactic features.

like *keep*, *leave*, or *prevent* are contrary to the semantics of CMCs. By removing large number of NON-CMC instances, we focus on how well the classifier performs on truly ambiguous cases. Furthermore, because these verbs have no instances of CMCs or DISPLACEs, only the negative label was reduced in size. Effectively, the removal of the verbs increases the proportion of the positive labels in the corpora. The numbers are shown in Table 6.

	CMC		DISPLACE	
Corpus	Before	After	Before	After
WSJ	14.8%	18.3%	8.86%	10.2%
WEB	29.2%	33.1%	24.2%	25.6%
BN	18.6%	21.6%	14.3%	15.7%

Table 6: Removed lemma count and effect on CMC label

Tables 7 and 8 show the precision and recall percentages and the f-score values when the instances of the most frequent NON-CMC verbs are removed from the training and testing data.

There is a general improvement in performance after the removal of the verbs from the data. The most marked improvement is in the WEB models (both CMC and DISPLACE) and the BN model's DISPLACE label classification. In particular the recall value shows improvement in these classifier models. As we have seen before, the **Baseline+PV** set and the full feature set show the best predictions. There is no noticeable improvement in the WSJ classifiers except for a slight (statistically insignificant) increase in the baseline values.

4.3 Random Downsampling of Negative Labels

As we have seen in Section 3.4, the CMC and the DISPLACE instances in WSJ are outnumbered by the negative, NON-CMC labels. The previous experiment on removing NON-CMC verbs effectively brought up the percentage of positive labels for the CMC and DISPLACE labels to 20% and 11%, respectively. However, label proportions of 20-80 or, worse, 11-89 are still highly unbalanced. Several studies have shown that in cases of training size imbalance, downsampling data can help with the performance of supervised classifiers (Weiss and Provost, 2001; Kubat and Matwin, 1997). Thus, for this experiment, we randomly downsample the negative labels in the WSJ training data to increase the percentage of positive labels[3]. For the sake of simplicity, we base the downsampling proportions on the CMC label: we cut the negative label so that the CMC label makes up 25% (Downsample1 ”D1”) and 30% (Downsample2 ”D2”)of the total data. The proportions of the DISPLACE labels are, therefore, 14.0% (D1)and 16.8% (D2), respectively.

Table 9 shows the performance of the WSJ models on the downsampled training set. The results indicate that the downsampling of the negative labels in the training data leads to increased performance. We have also tested the semantic feature set and the **Baseline+P** feature set as well. Their performances

[3] The downsampling was only applied to the training set, altering the distribution of labels only for the training data. The test set remains identical from its previous distribution in Section 4.2

	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Baseline	63.32	40.67	0.4953	69.0	54.7	0.611	75.7	60.0	0.669
Baseline+P	71.71	71.56	0.7164	80.7	86.8	0.836	79.2	81.4	0.803
Baseline+PV	70.97	73.33	0.7213	81.6	87.7	0.845	79.6	83.6	0.815
Semantic	69.37	68.44	0.6890	74.6	80.2	0.773	77.1	84.3	0.805
Full Set	73.88	76.67	0.7525	76.2	87.7	0.816	79.5	82.9	0.811

Table 7: System performance on CMC label classification with frequent NON-CMC verbs removed.

	WSJ			WEB			BN		
	P	R	F	P	R	F	P	R	F
Baseline	63.25	58.73	0.6091	70.3	63.4	0.667	71.1	57.8	0.638
Baseline+P	72.77	67.86	0.7023	74.1	76.8	0.754	79.4	75.5	0.774
Baseline+PV	74.89	69.84	0.7228	76.1	81.7	0.788	79.8	81.4	0.806
Semantic	71.81	64.68	0.6806	73.8	75.6	0.747	74.5	77.5	0.760
Full Set	73.60	73.02	0.7331	76.7	84.1	0.802	81.4	81.4	0.814

Table 8: System performance on DISPLACE label classification with frequent NON-CMC verbs removed.

are approximately equal with no significant difference from the **Baseline+PV**, so we do not include those numbers.

We observe a large increase in the recall values, resulting in the overall improvement of the classifiers trained on downsampled data[4] . Interestingly, with the random downsampling of the training data, we see a boost in the full feature set's performance far more than the **Baseline+PV** set's performance. In fact, in all cases we observed that the full features now show a significantly higher performance than the other features (McNemar, $p < 0.05$). The observed results for the two downsampled classifiers are not statistically distinct from one another.

CMC Classification:				
	D1		D2	
	R	F	R	F
Baseline	55.33	0.5900	68.00	0.6207
Baseline+PV	86.00	0.7866	89.11	0.7886
Full Set	88.89	0.8180	91.33	0.8171
DISPLACE Classification:				
	D1		D2	
	R	F	R	F
Baseline	69.05	0.6705	75.40	0.6798
Baseline+PV	85.32	0.7776	88.10	0.7776
Full Set	88.10	0.8177	91.27	0.8084

Table 9: Classification performance with downsampled training data.

5 Final Considerations and Future Work

We have presented our work on the automatic classification of CMCs in corpus data using the annotated data produced in our earlier study (Hwang et al., 2014). Our studies have shown that we can achieve the identification of caused motion instances at a higher rate than the inter-annotator agreement scores, the best performance that can be realistically expected. We have also shown that semantic information is highly indicative of the caused motion phenomenon, confirming our general intuition that the caused motion construction is a semantic phenomenon. We have also carried out cross-genre experiments, which we were not able to include in this paper in the interest of length. In these experiments, we find that syntax provides scalable features that generalize well across different types of text, producing better results in cross-genre experiments. We have also shown that the downsampling of the negative label has a positive impact on the classification of the labels.

[4] We only show the recall values in Table 9 as the increase observed in the f-score was mainly due to the recall values.

This work has made use of various gold annotations for the purposes of feature extraction. The most obvious next step in this investigation will involve experimentation with automatically obtained features. Additionally, we hope to examine the impact of further features. As the experiments have shown, the lexical and semantic features (lemma, preposition, VerbNet classes) surface as strong predictors of CMCs. It follows from this, that we should expand the feature search to other semantic information. One particular set of features that might be interesting, would be based on FrameNet frames. Since FrameNet's frames represent different conceptual semantic domains, features from FrameNet may be instrumental at capturing and highlighting the semantics of CMCs that are spread across VerbNet classes of differing semantic types. Moreover, it would also be interesting to expand on the lexical features: lexical features can be extended to not just the verb of the sentence but also to the noun phrases. Further investigation into using resources like WordNet (Miller, 1995; Fellbaum et al., 1998) might be needed to remedy sparse data issues that lexical features based on words from the noun phrases might create.

Acknowlegements

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also gratefully acknowledge the support of DARPA/IPTO funding under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, VN Supplement, and funding under the BOLT and Machine Reading programs. HR0011-11-C-0145 (BOLT) FA8750-09-C-0179 (M.R.).

References

Daniel M. Bikel, Richard Schwartz, and Ralph Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning: Special Issue on NL Learning*, 34.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Jinho Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. dissertation, University of Colorado at Boulder, Boulder, Colorado.

Cristiane Fellbaum, Joachim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantic annotation task. In Christiane Fellbaum, editor, *WordNet: An Electronic Database*. The MIT Press.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University Of Chicago Press.

Dan Guildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *40th Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July.

Jena D. Hwang, Rodney D. Nielsen, and Martha Palmer. 2010. Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8, Los Angeles, California, June. Association for Computational Linguistics.

Jena D. Hwang, Martha Palmer, and Annie Zaenen. 2013. Representing paths of motion in representing paths of motion in VerbNet. In Tracy Holloway King and Valeria de Paiva, editors, *From Quirky Case to Representing Space*. CSLI Online Publications.

Jena D. Hwang, Annie Zaenen, and Martha Palmer. 2014. Criteria for identifying and annotating caused motion constructions in corpus data. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: One-sided selection. *Proceedings of the Fourteenth International Conference on Machine Learning*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*, pages 114–119.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Vivek Srikumar. 2013. *Semantics of Role Labeling*. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel aati Hawwary, Mitchell Marcus, Ann Taylor, Eduard Hovy, Robert Belvin, and Ann Houston, 2012. *OntoNotes Release 4.99*, February.

Gary M. Weiss and Foster Provost. 2001. The effect of class distribution on classifier learning. Technical report, Rutgers University.