Class	Event Only (*n* = 0)			Event Only + Syntactic			Event + Syn + Hyper		
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
Short	0.742	0.465	0.571	0.758	0.587	0.662	0.707	0.606	0.653
Long	0.748	0.908	0.821	0.792	0.893	0.839	0.793	0.857	0.823
Overall Prec.	**74.7%**			**78.2%**			**76.6%**		
	Local Context (*n* = 2)			**Context + Syntactic**			**Context + Syn + Hyper**		
Short	0.672	0.568	0.615	0.710	0.600	0.650	0.707	0.606	0.653
Long	0.774	0.842	0.806	0.791	0.860	0.824	0.793	0.857	0.823
Overall Prec.	**74.2%**			**76.6%**			**76.6%**		

Table 7: Feature Evaluation with Different Feature Sets using SVM.

Class	Prec.	Rec.	F
Short	0.692	0.610	0.649
Long	0.779	0.835	0.806
Overall Prec.	**75.0%**		

Table 8: Test Performance on WSJ data.

P(A)	P(E)	Kappa
0.798	0.151	0.762
	0.143	0.764

Table 9: Inter-Annotator Agreement for Most Likely Temporal Unit.

data, and indicates the significant generalization capacity of the learned model.

5 Learning the Most Likely Temporal Unit

These encouraging results have prompted us to try to learn more fine-grained event duration information, viz., the most likely temporal units of event durations (cf. (Rieger 1974)'s ORDERHOURS, ORDERDAYS).

For each original event annotation, we can obtain the most likely (mean) duration by averaging its lower and upper bound durations, and assigning it to one of seven classes (i.e., second, minute, hour, day, week, month, and year) based on the temporal unit of its most likely duration.

However, human agreement on this more fine-grained task is low (44.4%). Based on this observation, instead of evaluating the *exact* agreement between annotators, an "*approximate* agreement" is computed for the most likely temporal unit of events. In "approximate agreement", temporal units are considered to match if they are the same temporal unit or an adjacent one. For example, "second" and "minute" match, but "minute" and "day" do not.

Some preliminary experiments have been conducted for learning this multi-classification task. The same data sets as in the binary classification task were used. The only difference is that the class for each instance is now labeled with one of the seven temporal unit classes.

Algorithm	Precision
Baseline	51.5%
C4.5	56.4%
NB	65.8%
SVM	**67.9%**
Human Agreement	79.8%

Table 10: Overall Test Precisions.

The baseline for this multi-classification task is always taking the temporal unit which with its two neighbors spans the greatest amount of data. Since the "week", "month", and "year" classes together take up largest portion (51.5%) of the data, the baseline is always taking the "month" class, where both "week" and "year" are also considered a match. Table 9 shows the inter-annotator agreement results for most likely temporal unit when using "approximate agreement". Human agreement (the upper bound) for this learning task increases from 44.4% to 79.8%.

10-fold cross validation was also used to train the learning models, which were then tested on the unseen held-out test set. The performance of the three algorithms is shown in Table 10. The best performing learning algorithm is again SVM with 67.9% test precision. Compared with the baseline (51.5%) and human agreement (79.8%), this again is a very promising result, especially for a multi-classification task with such limited training data. It is reasonable to expect that when more annotated data becomes available, the learning algorithm will achieve higher performance when learning this and more fine-grained event duration information.

Although the coarse-grained duration information may look too coarse to be useful, computers have no idea at all whether a meeting event takes seconds or centuries, so even coarse-grained estimates would give it a useful rough sense of how long each event may take. More fine-grained duration information is definitely more desirable for temporal reasoning tasks. But coarse-grained