Features	Original	Lemma	POS
Subject	presidents	president	NNS
Object	plan	plan	NN

Table 2: Syntactic relation features for the "signed" event in sentence (1).

Feature	1-hyper	2-hyper	3-hyper
Event	write	communicate	interact
Subject	corporate executive	executive	adminis-trator
Object	idea	content	cognition

Table 3: WordNet hypernym features for the event ("signed"), its subject ("presidents"), and its object ("plan") in sentence (1).

It is useful to extract the hypernyms not only for the event itself, but also for the subject and object of the event. For example, events related to a group of people or an organization usually last longer than those involving individuals, and the hypernyms can help distinguish such concepts. For example, "society" has a "group" hypernym (2 steps up in the hierarchy), and "school" has an "organization" hypernym (3 steps up). The direct hypernyms of nouns are always not general enough for such purpose, but a hypernym at too high a level can be too general to be useful. For our learning experiments, we extract the first 3 levels of hypernyms from WordNet.

Hypernyms are only extracted for the events and their subjects and objects, not for the local context words. For each level of hypernyms in the hierarchy, it's possible to have more than one hypernym, for example, "see" has two direct hypernyms, "perceive" and "comprehend". For a given word, it may also have more than one sense in WordNet. In such cases, as in (Gildea and Jurafsky, 2002), we only take the first sense of the word and the first hypernym listed for each level of the hierarchy. A word disambiguation module might improve the learning performance. But since the features we need are the hypernyms, not the word sense itself, even if the first word sense is not the correct one, its hypernyms can still be good enough in many cases. For example, in one news article, the word "controller" refers to an air traffic controller, which corresponds to the second sense in WordNet, but its first sense (business controller) has the same hypernym of "person" (3 levels up) as the second sense (direct hypernym). Since we take the first 3 levels of hypernyms, the correct hypernym is still extracted.

P(A)	P(E)	Kappa
0.877	0.528	0.740
	0.500	0.755

Table 4: Inter-Annotator Agreement for Binary Event Durations.

When there are less than 3 levels of hypernyms for a given word, its hypernym on the previous level is used. When there is no hypernym for a given word (e.g., "go"), the word itself will be used as its hypernyms. Since WordNet only provides hypernyms for nouns and verbs, "NULL" is used for the feature values for a word that is not a noun or a verb.

For the "signed" event in sentence (1), the extracted WordNet hypernym features for the event ("signed"), its subject ("presidents"), and its object ("plan") are shown in Table 3, and the feature vector is [write, communicate, interact, corporate_executive, executive, administrator, idea, content, cognition].

4 Experiments

The distribution of the means of the annotated durations in Figure 2 is bimodal, dividing the events into those that take less than a day and those that take more than a day. Thus, in our first machine learning experiment, we have tried to learn this *coarse-grained* event duration information as a binary classification task.

4.1 Inter-Annotator Agreement, Baseline, and Upper Bound

Before evaluating the performance of different learning algorithms, the inter-annotator agreement, the baseline and the upper bound for the learning task are assessed first.

Table 4 shows the inter-annotator agreement results among 3 annotators for binary event durations. The experiments were conducted on the same data sets as in (Pan et al., 2006). Two kappa values are reported with different ways of measuring expected agreement (P(E)), i.e., whether or not the annotators have prior knowledge of the global distribution of the task.

The human agreement before reading the guidelines (0.877) is a good estimate of the *upper bound* performance for this binary classification task. The *baseline* for the learning task is always taking the most probable class. Since 59.0% of the total data is "long" events, the baseline performance is 59.0%**.**