Sec	dep1	+hits	+V1	dep2	+hits	+V1	dep1-L	+hits-L	+V1-L	dep2-L	+hits-L	+V1-L
00	90.39	**90.94**	90.91	91.56	**92.16**	**92.16**	90.11	**90.69**	90.67	91.94	**92.47**	92.42
01	91.01	**91.60**	**91.60**	92.27	**92.89**	92.86	90.77	**91.39**	**91.39**	91.81	**92.38**	92.37
23	90.82	**91.46**	91.39	91.98	**92.64**	92.59	90.30	**90.98**	90.92	91.24	**91.83**	91.77
24	89.53	**90.15**	90.13	90.81	**91.44**	91.41	89.42	**90.03**	90.02	90.30	**90.91**	90.89

Table 3: Unlabeled accuracies (UAS) and labeled accuracies (LAS) on Section 0, 1, 23, 24. Abbreviation: dep1/dep2=first-order parser and second-order parser with the baseline features; +hits=N-gram features derived from the Google hits; +V1=N-gram features derived from the Google V1; suffix-L=labeled parser. Unlabeled parsers are scored using unlabeled parent predictions, and labeled parsers are scored using labeled parent predictions.

finding is that the N-gram features derived from Google hits are slightly better than Google V1 due to the large N-gram coverage, we will discuss later. As a final note, all the comparisons between the integration of N-gram features and the baseline features in Table 3 are mildly significant using the Z-test of Collins et al. (2005) ($p < 0.08$).

Type	Systems	UAS	CM
D	Yamada and Matsumoto (2003)	90.3	38.7
	McDonald et al. (2005)	90.9	37.5
	McDonald and Pereira (2006)	91.5	42.1
	Corston-Oliver et al. (2006)	90.9	37.5
	Hall et al. (2006)	89.4	36.4
	Wang et al. (2007)	89.2	34.4
	Carreras et al. (2008)	**93.5**	-
	GoldBerg and Elhadad (2010)†	91.32	40.41
	Ours	*92.64*	46.61
C	Nivre and McDonald (2008)†	92.12	44.37
	Martins et al. (2008)†	**92.87**	45.51
	Zhang and Clark (2008)	92.1	45.4
S	Koo et al. (2008)	93.16	-
	Suzuki et al. (2009)	**93.79**	-
	Chen et al. (2009)	93.16	47.15

Table 4: Comparison of our final results with other best-performing systems on the whole Section 23. Type D, C and S denote discriminative, combined and semi-supervised systems, respectively. † These papers were not directly reported the results on this data set, we implemented the experiments in this paper.

To put our results in perspective, we also compare them with other best-performing systems in Table 4. To facilitate comparisons with previous work, we only use Section 23 as the test data. The results show that our second order model incorporating the N-gram features (92.64) performs better than most previously reported discriminative systems trained on the Treebank. Carreras et al. (2008) reported a very high accuracy using information of constituent structure of TAG grammar formalism, while in our system, we do not use such knowledge. When compared to the combined systems, our system is better than Nivre and McDonald (2008) and Zhang and Clark (2008), but a slightly worse than Martins et al. (2008). We also compare our method with the semi-supervised approaches, the semi-supervised approaches achieved very high accuracies by leveraging on large unlabeled data directly into the systems for joint learning and decoding, while in our method, we only explore the N-gram features to further improve supervised dependency parsing performance.

Table 5 shows the details of some other N-gram sources, where **NEWS**: created from a large set of news articles including the Reuters and Gigword (Graff, 2003) corpora. For a given number of unique N-gram, using any of these sources does not have significant difference in Figure 3. Google hits is the largest N-gram data and shows the best performance. The other two are smaller ones, accuracies increase linearly with the log of the number of types in the auxiliary data set. Similar observations have been made by Pitler et al. (2010). We see that the relationship between accuracy and the number of N-gram is not monotonic for Google V1. The reason may be that Google V1 does not make detailed preprocessing, containing many mistakes in the corpus. Although Google hits is noisier, it has very much larger coverage of bigrams or trigrams.

Some previous studies also found a log-linear relationship between unlabeled data (Suzuki and Isozaki, 2008; Suzuki et al., 2009; Bergsma et al., 2010; Pitler et al., 2010). We have shown that this trend continues well for dependency parsing by using web-scale data (NEWS and Google V1).

[13]Google indexes about more than 8 billion pages and each contains about 1,000 words on average.