Model	Seg	POS
Hatori+12 SegTag	97.66	93.61
Hatori+12 SegTag(d)	98.18	94.08
Hatori+12 SegTagDep	97.73	94.46
Hatori+12 SegTagDep(d)	98.26	94.64
M. Zhang+14 EAG	97.76	94.36
Y. Zhang+15	98.04	94.47
SegTag(g)	98.41	**94.84**
SegTag	**98.60**	94.76

Table 5: Joint segmentation and POS tagging scores. Both scores are in F-measure. In Hatori et al. (2012), (d) denotes the use of dictionaries. (g) denotes greedy trained models. All scores for previous models are taken from Hatori et al. (2012), Zhang et al. (2014) and Zhang et al. (2015).

3.2 Results

3.2.1 Joint Segmentation and POS Tagging

First, we evaluate the joint segmentation and POS tagging model (SegTag). Table 5 compares the performance of segmentation and POS tagging using the CTB-5 dataset. We train two modles: a greedy-trained model and a model trained with beams of size 4. We compare our model to three previous approaches: Hatori et al. (2012), Zhang et al. (2014) and Zhang et al. (2015). Our SegTag joint model is superior to these previous models, including Hatori et al. (2012)'s model with rich dictionary information, in terms of both segmentation and POS tagging accuracy.

3.2.2 Joint Segmentation, POS Tagging and Dependency Parsing

Table 6 presents the results of our full joint model. We employ the greedy trained full joint model SegTagDep(g) and the beam decoding model SegTagDep. All scores for the existing models in this table are taken from Zhang et al. (2014). Though our model surpasses the previous best end-to-end joint models in terms of segmentation and POS tagging, the dependency score is slightly lower than the previous models. The greedy model SegTagDep(g) achieves slightly lower scores than beam models, although this model works considerably fast because it does not use beam decoding.

Model	Seg	POS	Dep
Hatori+12	97.75	94.33	81.56
M. Zhang+14 EAG	97.76	94.36	**81.70**
SegTagDep(g)	98.24	94.49	80.15
SegTagDep	**98.37**	**94.83**	81.42

Table 6: Joint Segmentation, POS Tagging and Dependency Parsing. Hatori et al. (2012)'s CTB-5 scores are reported in Zhang et al. (2014). EAG in Zhang et al. (2014) denotes the arc-eager model. (g) denotes greedy trained models.

Model	Seg	POS	Dep
Hatori+12	97.75	94.33	81.56
M. Zhang+14 STD	97.67	94.28	81.63
M. Zhang+14 EAG	97.76	94.36	81.70
Y. Zhang+15	98.04	94.47	82.01
SegTagDep(g)	98.24	94.49	80.15
SegTagDep	98.37	**94.83**‡	81.42‡
SegTag+Dep	**98.60**‡	94.76‡	**82.60**‡

Table 7: The SegTag+Dep model. Note that the model of Zhang et al. (2015) requires other base parsers. ‡ denotes that the improvement is statistically siginificant at $p < 0.01$ compared with SegTagDep(g) using paired t-test.

3.2.3 Pipeline of Our Joint SegTag and Dep Model

We use our joint SegTag model for the pipeline input of the Dep model (SegTag+Dep). Both SegTag and Dep models are trained and tested by the beam cost function with beams of size 4. Table 7 presents the results. Our SegTag+Dep model performs best in terms of the dependency and word segmentation. The SegTag+Dep model is better than the full joint model. This is because most segmentation errors of these models occur around named entities. Hatori et al. (2012)'s alignment step assumes the intra-word dependencies in words, while named entities do not always have them. For example, SegTag+Dep model treats named entity "海赛克", a company name, as one word, while the SegTagDep model divides this to "海" (sea) and "赛克", where "赛克" could be used for foreigner's name. For such words, SegTagDep prefers `SH` because `AP` has size-2 step of the character appending and intra-word dependency resolution, which does not exist for named entities. This problem could be solved by adding a special transition `AP_named_entity` which is similar to `AP` but with size-1 step and used