Better Chinese Word Segmentation using Dual Decomposition

Abstract

Cite (**?**).

1 Introduction

- CWS is essential for many downstream applications and its performance has direct impacts
- Two primary approaches: character-based and word-based
- Character based includes work like:
- Word based includes work like:
- In both cases out-of-vocabulary recall is a problem, but char-based sequence models win
- Solution has been proposed: mix models
- Existing work (Sun paper?) trains many models and uses voting
- We propose a simpler and more direct solution to model mixing: dual decomp to have two simultaneous mutually informative models

2 Methodology

- TODO: Work on this section together - explain dual decomp in this context in a simple way
- proposal: if Rob can understand the methodology then your average non-dual-decomp expert will understand it

3 Experiments

3.1 Accuracy

- show that this model does better than existing work
- biggest win is on R_{oov}

3.2 Efficiency

- show that this model is faster / simpler / etc than existing work
- important that methodology section is clear enough that other people could implement this

	Recall	Precision	F_1	OOV	R_{oov}	R_{iv}
CRF	0.978	0.969	0.973	0.035	0.723	0.987
PCT	0.978	0.971	0.974	0.035	0.730	0.987
DD	**0.981**	**0.973**	**0.977**	0.035	**0.741**	**0.989**

Table 1: Results on CTB-6 dataset.

4 Conclusion

[0] This result would place 2nd in the SIGHAN competition.

		Recall	Precision	F_1	OOV	R_{oov}	R_{iv}
SIGHAN 2005							
AS	SIGHAN winner	0.952	**0.951**	0.952	0.043	0.696	0.963
	CRF-Char	0.952	0.936	0.944	0.043	0.589	0.969
	Perceptron-Word	0.958	0.950	0.954	0.043	0.695	0.970
	Dual Decomp	0.959	0.949	**0.954**	0.043	0.677	**0.972**
PKU	SIGHAN winner	**0.953**	0.946	0.950	0.058	0.636	**0.972**
	CRF-Char	0.946	0.953	0.949	0.058	0.778	0.956
	Perceptron-Word	0.941	0.955	0.948	0.058	0.767	0.952
	Dual Decomp	0.948	**0.957**	**0.953**	0.058	**0.787**	0.958
CITYU	SIGHAN winner	0.941	**0.946**	0.943	0.074	0.698	0.961
	CRF-Char	0.947	0.940	0.943	0.074	**0.761**	0.962
	Perceptron-Word	0.943	0.940	0.942	0.074	0.717	0.961
	Dual Decomp	**0.950**	0.944	**0.947**	0.074	0.753	**0.965**
MSR	SIGHAN winner	0.962	0.966	0.964	0.026	0.717	0.968
	CRF-Char	0.964	0.966	0.965	0.026	0.713	0.971
	Perceptron-Word	0.970	0.972	0.971	0.026	0.746	0.976
	Dual Decomp	**0.973**	**0.974**	**0.974**	0.026	**0.760**	**0.979**
SIGHAN 2003							
AS	SIGHAN winner	0.966	0.956	0.961	0.022	0.364	**0.980**
	CRF-Char	0.969	0.969	0.969	0.022	0.748	0.974
	Perceptron-Word	0.967	0.967	0.967	0.022	0.729	0.972
	Dual Decomp	**0.970**	**0.971**	**0.971**	0.022	**0.775**	0.975
PKU	SIGHAN winner	**0.962**	0.940	0.951	0.069	0.724	**0.979**
	CRF-Char	0.954	0.952	0.953	0.069	0.803	0.965
	Perceptron-Word	0.949	0.952	0.950	0.069	0.790	0.960
	Dual Decomp	0.954	**0.954**	**0.954**	0.069	**0.806**	0.965
CITYU	SIGHAN winner	0.947	0.934	0.940	0.071	0.625	**0.972**
	CRF-Char	0.944	0.939	0.941	0.071	0.741	0.959
	Perceptron-Word	0.944	0.945	0.945	0.071	0.730	0.960
	Dual Decomp	**0.950**	**0.949**	**0.949**	0.071	**0.754**	0.965
CTB	SIGHAN winner	**0.886**	0.875	**0.881**	0.181	0.644	**0.927**
	CRF-Char	0.869	0.865	0.867	0.181	0.680	0.910
	Perceptron-Word	0.865	0.871	0.868	0.181	0.660	0.910
	Dual Decomp	0.876	**0.878**	0.877^0	0.181	**0.692**	0.917

Table 2: Results on SIGHAN 2005 and 2003 datasets.