Two Knives Cut Better Than One: Chinese Word Segmentation with Dual Decomposition

Abstract

There are two dominant approaches to the Chinese word segmentation problem: word-based and character-based models, each with respective strengths. Prior work has shown that gains in segmentation performance can be achieved from combining these two types of models; however, Past efforts have not provided a practical technique to allow mainstream adoption. We propose a method that effectively combines the strength of both segmentation schemes using an efficient dual-decomposition algorithm for joint inference. Our method is simple and easy to implement. Experiments on SIGHAN 2003 and 2005 evaluation datasets show that our method achieves the best reported results to date on 6 out of 7 datasets.

1 Introduction

Chinese text is written without delimiters between words; as a result, Chinese word segmentation (CWS) is an essential foundational step for many tasks in Chinese natural language processing. As demonstrated by [cite, cite, cite - Pichuan, what else?], the quality and consistency of segmentation has important downstream impacts on system performance in machine translation, [and...].

State-of-the-art performance in CWS is high, with F-scores in the upper 90s. Still, challenges remain. Unknown words, also known as out-of-vocabulary (OOV) words, lead to difficulties for word- or dictionary-based approaches. Ambiguity can cause errors when the appropriate segmentation is determined contextually, such as 才能 ("talent") and 才 / 能 ("just able") (Gao et al., 2003).

There are two primary classes of models: character-based (Xue, 2003; Tseng et al., 2005; Zhang et al., 2006; Wang et al., 2010) and word-based (Andrew, 2006; Zhang and Clark, 2007), with corresponding advantages and disadvantages. Sun (2010) details their theoretical distinctions: character-based approaches better model the internal compositional structure of words and are therefore more effective at inducing new out-of-vocabulary words; word-based approaches are better at reproducing the words of the training lexicon and can capture information from significantly larger contextual spans. Prior work has shown performance gains from combining these two types of models to exploit their respective strengths, but such approaches are often complex to implement and computationally expensive.

In this work, we propose a simple and principled joint decoding method for combining character-based and word-based segmenters based on dual decomposition. This method has strong optimality guarantees and works very well empirically. It is easy to implement and does not require retraining of existing character- and word-based segmenters. Experimental results on standard SIGHAN 2003 and 2005 bake-off evaluations show that our model outperforms the character and word baselines by a significant margin. In particular, it improves OOV recall rates and segmentation consistency, and gives the best reported results to date on 6 out of 7 datasets.

2 Models for CWS

In this section, we describe the character-based and word-based models we use as baselines, and review existing approaches to combine these models.

2.1 Character-based Models

In the most commonly used contemporary approach to character-based segmentation, first proposed by (Xue, 2003), CWS is seen as a character sequence tagging task, where each character is tagged on whether it is at the beginning, middle, or end of a

word. Conditional random fields (CRFs) are often used for this purpose (cite, cite, cite). In a CRF segmentation model, the probability of a label sequence is given by this equation:

Common linguistic features include character n-grams and morphological suffix/prefix features. Since these features capture information about the compositional properties of characters, they are likely to generalize well to unknown words.

2.2 Word-based Models

Initially word-based segmentation approaches employed simple heuristics like dictionary-lookup maximum matching (Chen and Liu, 1992), contemporary approaches use machine learning techniques to solve an argmax problem of the form:

The most successful such system reported to date is (Zhang and Clark, 2007)'s Perceptron-based model, which uses a search-based discriminative decoder to solve the ...

2.3 Mixing Models

Since the two types of models described above have different strengths and make different kinds of errors, various mixing approaches have been proposed to combine them (Wang et al., 2006; Lin, 2009; Sun et al., 2009; Sun, 2010; Wang et al., 2010).

(Lin, 2009) and (Wang et al., 2006) both incorporate a character-based Maxent local classifier model with a language model that captures more word-level context. In order to bring in a bigram language model, (2009) gave a heuristic decoding method that involves various forms of conditioning and back-off; (2006) gave a modified Viterbi algorithm with a complexity of $O(T^3)$.

(Sun et al., 2009) presented a latent-variable model which obtains good performance, but it is very complicated to implement and difficult to train. (Sun et al., 2009) use a method most similar to this work in that they also directly combine two segmenters – one char-based and one word-based. However, they do it through segmenter bagging, which requires training 50 or more individual segmenters and polling their results at test time.

These mixing models perform well on standard datasets, but are not in wide use because of their high computational costs and difficulty of implementation.

3 Dual-Decomposition

Dual decomposition offers a ideal framework for combining these two sources of signals without incurring high cost in model complexity (in contrast to (Sun et al., 2009)) or decoding efficiency (in contrast to bagging in (Wang et al., 2006; Sun, 2010)). DD has been successfully applied to similar situations where we want to combine local model with global models, for example, in dependency parsing (Koo et al., 2010)), bilingual sequence tagging (Wang et al., 2013) and word alignment (). Give a brief description of DD algorithm, focus on the intuition. See (Wang et al., 2013) and (DeNero and Macherey, 2011) for a good short introduction example. Refer users to (Rush and Collins, 2012) for a full tutorial on dual decomp. The modification to Viterbi decoding is exactly the same as in (Wang et al., 2013) and (DeNero and Macherey, 2011). The modification to the beam-search is similar, each time we extend a hypothesis with a new character, depending if the new character is appended to the last word or starting a new word, the corresponding DD penalty is factored into the score for the new hypothesis.

4 Experiments

In this work, we employ two baseline models — a character-based CRF and a word-based perceptron — and test the performance of jointly decoding these baseline systems with dual decomposition.

For our character-based CRF, we use the open-source Stanford CRF segmenter described in (Tseng et al., 2005).[1] In this system we use L2 regularization with a value of X and sigma set to 3.

For our word-based perceptron, we use a reimplementation of (Zhang and Clark, 2007), run with 10 iterations of training.

For joint decoding with dual decomposition, we use an initial step size set to 0.1, and run for 100 iterations.

We use the same development set employed by (Zhang and Clark, 2007), with Chinese Treebank sections 1-270, 400-931, and 1001-115 used as training data and sections 271-300 used as development data for tuning the hyperparameters of all three systems.

[1] http://nlp.stanford.edu/software/segmenter.shtml

SIGHAN 2005				
	AS	PU	CU	MSR
Best 05	95.2	95.0	94.3	96.4
Zhang et al. 06	94.7	94.5	94.6	96.4
Z&C 07	94.6	94.5	95.1	97.2
Sun et al. 09	-	95.2	94.6	97.3
Sun 10	95.2	95.2	**95.6**	96.9
Dual-decomp	**95.4**	**95.3**	94.7	**97.4**
SIGHAN 2003				
Best 03	96.1	95.1	94.0	
Peng et al. 04	95.6	94.1	92.8	
Z&C 07	96.5	94.0	94.6	
Dual-decomp	**97.1**	**95.4**	**94.9**	

Table 2: Performance of dual-decomposition (DD) in comparison to past published results on SIGHAN 2003 and 2005 datasets. Best reported F_1 score for each dataset is highlighted in bold. *Z&C 07* refers to Zhang and Clark (2007). *Best 03, 05* are results of the winning systems for each dataset in the respective shared tasks.

4.1 Datasets

We run our experiments on the SIGHAN 2003 (Sproat and Emerson, 2003) and 2005 (Emerson, 2005) bake-off datasets. Specifically, we use the standard training and test splits for the 2003 Academia Sinica (AS), Peking University (PU), and City University of Hong Kong (CU) datasets as well as the 2005 AS, PU, CU, and Microsoft Research (MSR) datasets. We do not use the 2003 Chinese Treebank dataset because... [why exactly?]

5 Results

Table 1 shows our empirical results on each of the 7 datasets from SIGHAN 2003 and 2005. For each dataset, the top three lines show the performance of our character-based CRF and word-based Perceptron (PCPT) systems, as well as their performance under joint decoding with dual decomposition (DD), and the bottom five lines allow for comparison with available results from prior and related work.

Dual decomposition outperforms our baselines in terms of F_1 on all seven datasets, demonstrating a robustness across domains and segmentation standards regardless of which baseline model was stronger. Of particular note is DD's out-of-vocabulary recall (R_{oov}), which outperforms our baselines on 5 out of 7 datasets and outperforms all existing available results in this category on every dataset.

6 Discussion

On the whole, dual decomposition produces state-of-the-art segmentations that are more accurate, more consistent, and more successful at inducing out-of-vocabulary words than the baseline systems that it combines.

6.1 Dual decomposition convergence

Plot the histogram of the number of iterations it takes to converge, and percentage of optimality.

6.2 Consistency

(Chang et al., 2008) have shown that increased segmentation consistency is correlated with better machine translation performance. Following their method for calculating the conditional entropy of a segmentation system, we see in Table [insert table] that our dual decomposition method achieves the most consistent segmentations (lowest conditional entropy) on 6 out of 7 datasets.

6.3 Error analysis

Since dual decomposition is a method of joint decoding, it is liable to reproduce errors made by the constituent systems. In the example below, dual decomposition output follows the incorrect segmentation of the character-based CRF in oversegmenting the compound "sea, land, and air."

Gloss	Large-scale / sea, land, and air / joint / military exercises
Gold	大规模 / 海陆空 / 联合 / 军演
CRF	大规模 / 海 / 陆空 / 联合 / 军演
PCPT	大规模 / 海陆空 / 联合 / 军演
DD	大规模 / 海 / 陆空 / 联合 / 军演

Nevertheless, in many cases the relative confidence of each model means that dual decomposition is capable of using information from both sources to generate a series of correct segmentations better than either baseline model alone. The example below shows a difficult-to-segment proper name comprised of common characters, which results in undersegmentation by the character-based CRF and oversegmentation by the word-based Perceptron, but our method achieves the correct middle ground.

Gloss	Tian Yage / 's / creations
Gold	田雅各 / 的 / 创作

	AS					PU				
	R	P	F_1	R_{oov}	C_{onst}	R	P	F_1	R_{oov}	C_{onst}
Char-based CRF	95.2	93.6	94.4	58.9	0.064	94.6	95.3	94.9	77.8	0.089
Word-based Perceptron	95.8	95.0	95.4	69.5	0.060	94.1	95.5	94.8	76.7	0.099
Dual-decomp	95.9	94.9	95.4	67.7	0.055	94.8	95.7	95.3	78.7	0.086
	CU					MSR				
	R	P	F_1	R_{oov}	C_{onst}	R	P	F_1	R_{oov}	C_{onst}
Char-based CRF	94.7	94.0	94.3	76.1	0.065	96.4	96.6	96.5	71.3	0.074
Word-based Perceptron	94.3	94.0	94.2	71.7	0.073	97.0	97.2	97.1	74.6	0.063
Dual-decomp	95.0	94.4	94.7	75.3	0.062	97.3	97.4	97.4	76.0	0.055

Table 1: Results on SIGHAN 2005 datasets. R_{oov} denotes OOV recall, and C_{onst} denotes segmentation consistency.

CRF 田雅各的 / 创作
PCPT 田雅 / 各 / 的 / 创作
DD 田雅各 / 的 / 创作

A powerful feature of the dual decomposition approach is that it can generate correct segmentation decisions in cases where a voting or polling-of-experts model could not, since joint decoding allows the sharing of information at decoding time. In the following example, both baseline models miss the contextually clear grammatical role of 上 ("above") and instead produce the otherwise common compound 上去 ("go up"); dual decomposition allows the model to generate the correct segmentation.

English Enjoy / a bit of / snack food / , ...
Gold 享受 / 一点 / 点心 / ,
CRF 享受 / 一点点 / 心 / ,
PCPT 享受 / 一点点 / 心 / ,
DD 享受 / 一点 / 点心 / ,

We found more than 400 such surprisingly accurate instances in our dual decomposition output.

7 Conclusion

In this paper we presented an approach to Chinese word segmentation using dual decomposition for system combination. We demonstrated that this method allows for joint decoding of existing CWS systems that performs better than either system alone, and further achieves the best performance reported to date on standard datasets for the task.

We also demonstrated that our joint system produces more consistent segmentations than our baselines. This property that has potentially important downstream impacts for many Chinese NLP tasks, though it remains to test this empirically.

Perhaps most importantly, our approach is straightforward to implement and does not require retraining of the underlying segmentation models used. This suggests its potential for broader applicability in real-world settings than existing approaches to combining character-based and word-based models for Chinese word segmentation.

References

Galen Andrew. 2006. A hybrid Markov/semi-Markov conditional random field for sequence segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pichuan Chang, Michel Galley, and Chris Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the ACL Workshop on Statistical Machine Translation*.

Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for mandarin chinese sentences. In *Proceedings of the 14th Conference on Computational Linguistics*.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Jianfeng Gao, Mu Li, and Chang-Ning Huang. 2003. Improved source-channel models for Chinese word segmentation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, and David Sontag. 2010. Dual decomposition for parsing with non-projective head automata. In *Proceedings of EMNLP*.

Dekang Lin. 2009. Combining language modeling and discriminative classification for word segmentation. In *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Alexander M. Rush and Michael Collins. 2012. A tutorial on dual decomposition and Lagrangian relaxation for inference in natural language processing. *JAIR*, 45:305–362.

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language Processing*.

Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Weiwei Sun. 2010. Word-based and character-basedword segmentation models: Comparison and combination. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING)*.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurasfky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian, and Xihong Wu. 2006. Chinese word segmentation with maximum entropy and n-gram language model. In *Proceedings of the fifth SIGHAN workshop on Chinese language Processing*.

Kun Wang, Chengqing Zong, and Keh-Yih Su. 2010. A character-based joint model for chinese word segmentation. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING)*.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, pages 29–48.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for Chinese word segmentation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL)*.