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

Author 1
XYZ Company
111 Anywhere Street
Mytown, NY 10000, USA
author1@xyz.org

Author 2
ABC University
900 Main Street
Ourcity, PQ, Canada A1A 1T2
author2@abc.ca

Abstract

There are two dominant approaches to the Chinese word segmentation problem: word-based and character-based models, each with respective strengths and weaknesses. Prior work has shown that gains in segmentation performance can be achieved from combining these two types of models; however, past efforts involve either training many instances of segmenters and polling their results, or designing complex latent-variable methods that incur heavy computational costs. In this paper, we propose an effective and consistent joint decoding method using dual decomposition, which does not require any additional training. Our method is simple and easy to implement, and achieves the best reported results to date on 6 out of 7 standard SIGHAN evaluation 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”) (**?**).

There are two primary classes of models: character-based (**?**; **?**; **?**; **?**) and word-based (**?**; **?**), with corresponding advantages and disadvantages. (**?**) 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 (?), 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 (?), contemporary approaches use machine learning techniques to solve an argmax problem of the form:

The most successful such system reported to date is (?)'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 (?; ?; ?; ?; ?).

(?) and (?) 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, (?) gave a heuristic decoding method that involves various forms of conditioning and back-off; (?) gave a modified Viterbi algorithm with a complexity of $O(T^3)$.

(?) presented a latent-variable model which obtains good performance, but it is very complicated to implement and difficult to train. (?) 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 (?)) or decoding efficiency (in contrast to bagging in (?; ?)). DD has been successfully applied to similar situations where we want to combine local model with global models, for example, in dependency parsing (?)), bilingual sequence tagging (?) and word alignment (). Give a brief description of DD algorithm, focus on the intuition. See (?) and (?) for a good short introduction example. Refer users to (?) for a full tutorial on dual decomp. The modification to Viterbi decoding is exactly the same as in (?) and (?). 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 (?).[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 (?), 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 (?), with Chinese Treebank sections 1-270, 400-931,

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

and 1001-115 used as training data and sections 271-300 used as development data for tuning the hyper-parameters of all three systems.

4.1 Datasets

We run our experiments on the SIGHAN 2003 (**?**) and 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

(**?**) have shown that increased segmentation consistency is correlated with better machine translation performance. Following their method for calculating the conditional entropy of a segmentation

SIGHAN 2005

		F_1	R_{oov}	R_{iv}
AS	CRF	94.4	58.9	96.9
	PCPT	95.4	69.5	97.0
	DD	95.4	67.7	97.2
	Best 05	95.2	69.6	96.3
	Zhang et al. 06	94.7	-	-
	Z&C 07	94.6	-	-
	Sun et al. 09	-	-	-
	Sun 10	95.2	-	-
PU	CRF	94.9	77.8	95.6
	PCPT	94.8	76.7	95.2
	DD	95.3	78.7	95.8
	Best 05	95.0	63.6	97.2
	Zhang et al. 06	94.5	-	-
	Z&C 07	94.5	-	-
	Sun et al. 09	95.2	77.8	-
	Sun 10	95.2	-	-
CU	CRF	94.3	76.1	96.2
	PCPT	94.2	71.7	96.1
	DD	94.7	75.3	96.5
	Best 05	94.3	69.8	96.1
	Zhang et al. 06	94.6	-	-
	Z&C 07	95.1	-	-
	Sun et al. 09	94.6	68.8	-
	Sun 10	95.6	-	-
MS	CRF	96.5	71.3	97.1
	PCPT	97.1	74.6	97.6
	DD	97.4	76.0	97.9
	Best 05	96.4	71.7	96.8
	Zhang et al. 06	96.4	-	-
	Z&C 07	97.2	-	-
	Sun et al. 09	97.3	72.2	-
	Sun 10	96.9	-	-

SIGHAN 2003

AS	CRF	96.9	74.8	97.4
	PCPT	96.7	72.9	97.2
	DD	97.1	77.5	97.5
	Best 03	96.1	36.4	98.0
	Peng et al. 04	95.6	-	-
	Z&C 07	96.5	-	-
PU	CRF	95.3	80.3	96.5
	PCPT	95.0	79.0	96.0
	DD	95.4	80.6	96.5
	Best 03	95.1	72.4	97.9
	Peng et al. 04	94.1	-	-
	Z&C 07	94.0	-	-
CU	CRF	94.1	74.1	95.9
	PCPT	94.5	73.0	96.0
	DD	94.9	75.4	96.5
	Best 03	94.0	62.5	97.2
	Peng et al. 04	92.8	-	-
	Z&C 07	94.6	-	-

Table 1: Results on SIGHAN 2005 and 2003 datasets.

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.

Consistency Results, Sighan 2003:

as crf 0.037814800237 as pct 0.038997075095 as dd 0.0372484903836 cityu crf 0.0923886605473 cityu pct 0.097496193366 cityu dd 0.0843064189868 ctb crf 0.170532711404 ctb pct 0.181993909505 ctb dd 0.161722439818 pku crf 0.0429803512049 pku pct 0.0552572297562 pku dd 0.0460509973844

Consistency Results, Sighan 2005:

as crf 0.0635314146511 as pct 0.0603263902622 as dd 0.055097582491 cityu crf 0.0649537861096 cityu pct 0.0730776631941 cityu dd 0.0618197636595 msr crf 0.073782773985 msr pct 0.0628241148726 msr dd 0.054883418641 pku crf 0.0888607888366 pku pct 0.0996634383663 pku dd 0.086635535204

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.”

English	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.

English	Tian Yage’s creations
Gold	田雅各 / 的 / 创作
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.