How unsupervised learning affects character tagging based Chinese Word Segmentation: A quantitative investigation

Article · July 2009

DOI: 10.1109/ICMLC.2009.5212769

Ruifeng Xu
Harbin Institute of Technolog , Shenzhen Graduate School
240 PUBLICATIONS **4,510** CITATIONS

SEE PROFILE

Hai Zhao
Northeastern University (Shenyang, China)
494 PUBLICATIONS **6,214** CITATIONS

SEE PROFILE

How Unsupervised Learning Affects Character Tagging based Chinese Word Segmentation: A Quantitative Investigation

YAN SONG CHUNYU KIT RUIFENG XU HAI ZHAO

Department of Chinese, Translation and Linguistics
City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong
E-MAIL: {yansong, ctckit, ruifenxu, haizhao}@cityu.edu.hk

Abstract

Integrating global information of unsupervised segmentation into Conditional Random Fields (CRF) learning has been proved effective to enhance the performance of the character tagging based Chinese Word Segmentation. By comparing CRF models with and without unsupervised learning enhancement, we investigate how unsupervised learning affects the performance. Especially, two kinds of segmented words, in-vocabulary and out-of-vocabulary words, are separately analyzed case by case to see what part of those words are affected by unsupervised learning. In addition, the cost of the additional features derived from unsupervised segmentation are also taken into account and evaluated.

Keywords:

Unsupervised learning; Chinese word segmentation; in-vocabulary words; out-of-vocabulary words; frequent substring extraction

1. Introduction

Character tagging was well applied in Chinese Word Segmentation (CWS) in recent years [1][2][3][9][10]. The essential of the approach is to assign a pre-defined in-word-position tag to a given character from an input sequence, then restore word boundaries according to these output tags. For instance, in a four-tag set, B, I, E, S [1, 4], which stand for the beginning of a word, the middle of a word, the end of a word and a single character word, respectively, are used for describing the role of a character in a sequence. With a proper learning framework, such as CRF [5], a character tagging approach may bring state-of-the-art performance.

To further improve the CWS performance, some additional features have been developed for further performance improving.So far the most significant enhancement for CWS was conducted by the features from unsupervised learning as in [9]. According to Table 1 cited from [9], a report for SIGHAN Bakeoff 4 [7] CWS task, the results generated with the assistance of an unsupervised learning criterion, AV (Access Variety) [8] is much better than those from the basic CRF tagger (which is referred to the model without using unsupervised technique.). Especially, the performance related to out-of-vocabulary (OOV) words were also significantly improved.

To better understand what and how the unsupervised segmentation works, we quantitatively analyze the CWS results on two corpora from SIGHAN Bakeoff-3 [6]. The comparisons between the results clearly depict the enhanced model (referred to the model with unsupervised learning enhancement) using its superiority, as well as the limitation. We then discuss about the principles and characteristics of the assistance that unsupervised learning helps the character tagging based CWS.

2. Two Taggers

For the baseline CRF tagger, we adopt the system described in [9], in which 6 tags, namely, B, B_2, B_3, M, E and S, are used for tagging. Their functions are illustrated in Table 2.

Six n-gram feature templates, C_{-1}, C_0, C_1, $\mathrm{C}_{-1}\mathrm{C}_0$, $\mathrm{C}_0\mathrm{C}_1$ and $\mathrm{C}_{-1}\mathrm{C}_1$, are applied in both basic and enhanced CRF tagger. The subscripts, -1, 0, 1, denote the previous, current and successor position in the sequence, respectively.

The enhanced CRF tagger is adopted from the baseline one by integrating with the information of unsupervised learning. We use frequent substring extraction (FSE) as our unsupervised learning criterion. Among those CRF taggers

[1] CWS performance is measured by F-score, F=$2RP/(R+P)$, where R and P are recall and precision respectively.

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

Table 1. Comparisons of performances from different CRF taggers

Features	Data	F[1]	P	R	F_{IV}	P_{IV}	R_{IV}	F_{OOV}	P_{OOV}	R_{OOV}
N-gram	CityU[2]	0.9426	0.9410	0.9441	0.9640	0.9636	0.9645	0.7063	0.6960	0.7168
	CKIP	0.9421	0.9387	0.9454	0.9607	0.9581	0.9633	0.7113	0.7013	0.7216
	CTB	0.9634	0.9641	0.9627	0.9738	0.9761	0.9715	0.7924	0.7719	0.8141
	NCC	0.9333	0.9356	0.9311	0.9536	0.9612	0.9461	0.5678	0.5182	0.6280
	SXU	0.9552	0.9559	0.9544	0.9721	0.9767	0.9675	0.6640	0.6223	0.7116
N-gram + AV	CityU	0.9510	0.9493	0.9526	0.9667	0.9626	0.9708	0.7698	0.7912	0.7495
	CKIP	0.9470	0.9440	0.9501	0.9623	0.9577	0.9669	0.7524	0.7649	0.7404
	CTB	0.9470	0.9440	0.9501	0.9697	0.9704	0.9691	0.7745	0.7761	0.7730
	NCC	0.9470	0.9440	0.9501	0.9573	0.9583	0.9562	0.6080	0.5984	0.6179
	SXU	0.9470	0.9440	0.9501	0.9752	0.9764	0.9740	0.7292	0.7159	0.7429

Table 2. Six tags for character tagging

Word Length	Word Tagging
1	S
2	BE
3	BB_2E
4	BB_2B_3E
5	BB_2B_3ME
$\geq$6	$BB_2B_3M...ME$

Table 3. Statistics of the experiment data

Source	Training Words	Testing Words	OOV Rate
CityU	1.6M	220K	0.040
UPUC	509K	155K	0.088

enhanced by different unsupervised learning approaches, [11] proved that FSE could help output a competitive performance[2] in contrast to AV, which shows its effectiveness in Table 1. Thus CWS results generated by FSE enhanced CRF model is capable to support our investigation as well as that by AV enhanced CRF model. Another reason we use FSE is that it is an easily-implemented and straightforward unsupervised substring exrtaction method, tokens derived by FSE are usually more than those by other criteria (many tokens extracted by FSE may not be valid AV tokens), thus more information could be exploited for enhancing the basic CRF tagger. For those unsupervised learning information, the extracted tokens are assumed as the word candidates plus 5 unigram feature templates to represent the role of current character in those tokens, and the tackling follows the way described in [9].

3. Analysis

3.1. Data

We use two SIGHAN Bakeoff-3 corpora, CityU (City University of Hong Kong) and UPUC (University of Pennsylvania/University of Colorado), for our empirical study. The statistics of the training and testing sets is in Table 3. Note the OOV rate indicates the difficulty of the task to some extent, since the more OOV words in the testing data, the more difficult it is to be correctly segmented.

3.2. Overall Performances

For the CWS process, we follow the constraints by the closed track of SIGHAN Bakeoffs. Table 4 shows the overall performance of the basic and enhanced CRF tagger on two selected corpora.

3.3. Statistics from IV words

It is shown in the Table 4 that the precision and recall are all improved on both IV and OOV words. Consider that our unsupervised learning is based on the frequency of the substrings, a word-frequency criterion is proposed for our analysis. Words segmented by different CRF taggers with different frequencies are compared with the gold standard set. We divide the words into 5 categories by their occurrences, from 1 to 5, and those words appears in the corpus more than 5 times are also placed to the 5th category since it is found that high frequency words are easy to be extracted out and the variance of the CWS performance is often caused by those low-frequent words. The result of numbers and percentage of those words is presented in Table 5. Note that in our analysis we only counter the number of tokens but not the word types.

[2] Actually, [11] used the FSE after reduction in its work, but our experimental results show FSE without reduction could obtain better performance.

Table 4. Overall performance on CityU and UPUC tracks

Features	Data	F	P	R	F$_{IV}$	P$_{IV}$	R$_{IV}$	F$_{OOV}$	P$_{OOV}$	R$_{OOV}$
N-gram	CityU	0.9692	0.9695	0.9689	0.9796	0.9824	0.9767	0.7380	0.6969	0.7843
	UPUC	0.9324	0.9264	0.9386	0.9534	0.9466	0.9603	0.7122	0.7127	0.7117
N-gram + FSE	CityU	0.9733	0.9734	0.9732	0.9812	0.9821	0.9804	0.7877	0.7738	0.8022
	UPUC	0.9427	0.9384	0.9470	0.9583	0.9522	0.9645	0.7762	0.7881	0.7646

Table 5. Statistics of the IV words with different frequencies

		1	2	3	4	≥5
CityU	St	6924	5940	5151	4276	188996
	N	5673 (81.9%)	5392 (90.8%)	4848 (94.1%)	4060 (94.9%)	187891 (99.4%)
	N+F	5955 (86.0%)	5508 (92.7%)	4908 (95.3%)	4092 (95.7%)	188081 (99.5%)
UPUC	St	3410	3264	2886	2524	128228
	N	2515 (73.8%)	2732 (83.7%)	2541 (88.0%)	2292 (90.8%)	127452 (99.4%)
	N+F	2733 (80.1%)	2828 (86.6%)	2592 (89.8%)	2356 (93.3%)	127742 (99.6%)

In Table 5, St denotes the standard corpus, and N, N+F stand for the basic N-gram CRF model's output and the output of N-gram CRF model with FSE features, respectively. We can easily read from the data that correct IV words generated by N+F model are more than that from the other in all categories in both CityU and UPUC corpora. Specifically, more words which occurred only once (frequency=1) are correctly extracted by N+F model. It is proved in CWS results that the unsupervised learning helps the CRF tagger accurately find out the low frequency words which are ambiguous to be segmented, since the corrected words are precisely all wrongly segmented ones by the basic CRF tagger, and none of those correctly segmented words are affected.

Besides the frequency criterion, another key factor that affects the word identification is word length. Since the primary n-gram feature used in CRF tagger is about character position, the length of the word directly affects the tag to be assigned to a character. Similar to the statistics on the words with different frequencies, a statistics is conducted on the correct words with different lengths. Five categories with length=1, 2, 3, 4 and ≥5 are taken into account. Note the reason for this 5-category classification is different from the frequency statistics, as in the processing described in [9] that we follow in our experiments, the features from unsupervised learning are focused on those word candidates which are no longer than 5 characters in our training. This statistic results can evaluate the effectiveness of N+F CRF model more accurately. The feature template from unsupervised learning well fit the CRF training of the words with different length. For those correctly extracted words, it is shown that the shorter they are, the harder they can be segmented. The percentage of the missing words in segmentation that is from 2.4% in category 1 to 0 in category 5 in UPUC corpus as shown in Table 6 indicates that the greatest loss is from the once-appeared words.

However, the overall improvement by IV words is still limited due to the huge amount of IV words. For instance, over 700 words correctly segmented could only contribute 0.5% on F-score improvement in UPUC corpus. As for CityU corpus, total 210054 IV words are segmented by N model while 210934 IV words segmented by N+F model, although more correct words are generated, more incorrect words are also extracted, they bring us a lower precision (0.03%) as a result. Since CWS is concerned with both IVs and OOVs, the improvement on finite set of IVs can not solve the wrongly identified boundary caused by OOVs, thus we will discuss the case of OOV in the next part.

3.4. Statistics from OOV words

OOV is one of two major factors which affects the performance of a CWS system [12]. As it is seen from Table 4, the F-score on OOV words improves greatly. In contrast to the result from IV words, the OOV words extracted by the enhanced model contribute at least 7% improvement in precision and 2% in recall rate, which are significant better than the basic model.

We also use the 5-category system on frequency and length to investigate the quantitative difference between the OOV words generated by two models, see Table 7 and 8. For those numbers in Table 7, it is seen that correct words segmented by N+F model are more than those segmented by N model in both CityU and UPUC test when they appear more than once in the corpus. The unsupervised learning could help better dealing with words appearing many times in the data. Since the more times a word occurs in

Table 6. Statistics of the IV words with different lengths

		1	2	3	4	≥5
CityU	St	100597	97698	10137	2590	265
	N	98477 (97.9%)	96591 (98.9%)	9992 (98.6%)	2546 (98.3%)	258 (97.4%)
	N+F	98992 (98.4%)	96754 (99.0%)	9997 (98.6%)	2540 (98.1%)	261 (98.5%)
UPUC	St	69127	65876	5377	702	230
	N	67069 (97.0%)	64448 (97.8%)	5113 (95.1%)	673 (95.9%)	229 (99.6%)
	N+F	67461 (97.6%)	64684 (98.2%)	5182 (96.4%)	694 (98.9%)	230 (100%)

Table 7. Statistics of the OOV words with different frequencies

		1	2	3	4	≥5
CityU	St	3941	1414	774	452	2306
	N	2836 (72.0%)	1130 (79.9%)	648 (83.7%)	372 (82.3%)	1995 (86.5%)
	N+F	2720 (69.0%)	1172 (82.9%)	672 (86.8%)	392 (86.7%)	2181 (94.6%)
UPUC	St	5049	2102	1230	716	4455
	N	3390 (67.1%)	1602 (76.2%)	939 (76.3%)	520 (72.6%)	3218 (72.2%)
	N+F	3289 (65.1%)	1676 (79.7%)	999 (81.2%)	608 (84.9%)	3835 (86.1%)

a corpus, the higher possibility it could be extracted out, the features from unsupervised learning reinforce the wordhood of those high frequent substrings, make them to be easily segmented as words. In contrast to the improvement of higher frequent OOVs, the unsupervised learning also affects the extraction of lower frequent OOVs, as in both CityU and UPUC corpora, the numbers of correctly segmented OOVs which appeared only once are all decreased.

Different from the situation of frequencies, in statistics on length we find that those words with length equal to 2 are the worse part while others are improved. Like the situation in IV words, the investigation conducted to unveil the details show us the decreased part are contributed by those once-occurred words. For instance, in total 5049 OOV words that appear only once in UPUC test data, there are 2508 two-character words. And 2034 out of these 2497 OOVs are extracted by the basic model in while only 1881 of those are extracted by the enhanced model. But this case is changed when we turn to 634 once-occurred OOVs with four-character, 203 extracted by the basic model and 219 by the enhanced model. Also the more characters in a word candidate, the higher probability it could be recognized as a word by FSE method whose target is frequent strings.

4. Discussion

According to the above analysis, we may conclude what part of the CWS results are improved by the unsupervised enhancement in a supervised learning framework. Actually, there is other means to make use of such guidance. [13] also used an unsupervised learning approach, left context entropy [14], as a post processing step to improve the OOV performance. The results show that strategy for CWS is not so efficient as integrating them into a unified learning process. In brief, the reason why supervised learning benefits from such features is because the unsupervised information acts as an auxiliary guidance, not a decisive rule to segment words, while the former is a feasible parameter and the latter always plays as a binary judgment [15]. Let those parameters to be trained in an effective learning model, CRF, is superb than turning them to fixed rules in post process.

It is worth pointing out that OOVs and IVs are two relative parts in the CWS result. Usually more correctly extracted OOVs could generate more accurate word boundaries, which can make the rest IVs to be also correctly recognized. Let's see a statistics on a macro criterion, the rate of incorrect segmented sentence, which defines, if a sentence contains more than one segmentation error, the sentence is an incorrectly segmented one. In CityU corpus, there are 3527 wrongly segmented sentences generated by the basic model and 3142 generated by the enhanced model; while in UPUC corpus, there are 4964 wrongly segmented sentences generated by the basic model and 4439 generated by the enhanced model, nearly 400 and over 500 sentences are corrected in each test due to the help of unsupervised learning. To penetrate into the details, we find in UPUC test data, over 95% OOVs are adjacent to IVs. Once an OOV is correctly extracted, two IVs (if the OOV is between them) are consequently properly segmented. The above analysis could reasonably explain the relative performance improvement on IVs as more OOVs are correctly extracted. Another factor to be mentioned here is all extracted OOVs by

Table 8. Statistics of the OOV words with different lengths

		1	2	3	4	≥5
CityU	St	86	3379	3817	1002	603
	N	32 (37.2%)	2916 (86.3%)	3024 (79.2%)	571 (57.0%)	438 (72.6%)
	N+F	40 (46.5%)	2812 (83.2%)	3183 (83.4%)	632 (63.1%)	470 (77.9%)
UPUC	St	169	7025	4765	1086	507
	N	59 (34.9%)	6001 (85.4%)	2938 (61.7%)	350 (32.2%)	321 (63.3%)
	N+F	67 (39.6%)	5897 (83.9%)	3620 (76.0%)	458 (42.2%)	365 (72.0%)

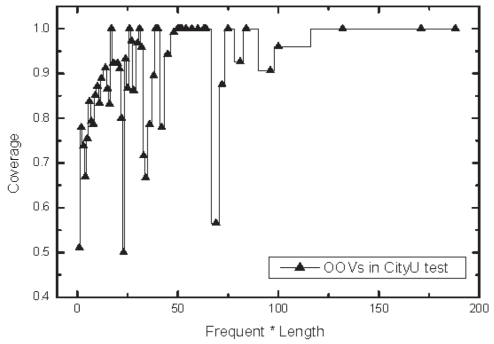

Figure 1. The coverage of OOVs in CityU track.

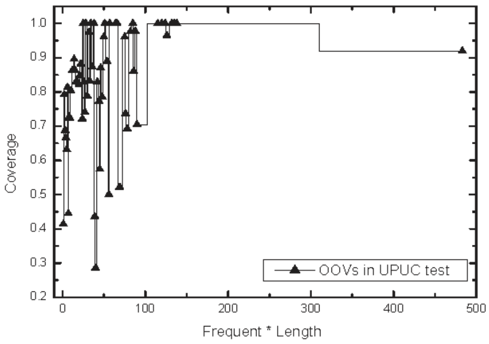

Figure 2. The coverage of OOVs in UPUC track.

the enhanced model are less than those by the basic model, as they concentrate on the higher frequent words. As a result, the correct rate of OOVs increases, the precision is thus improved significantly than the recall.

We also observe the consequence caused by both frequency f and length l, as in low frequent words, the longer it be, the more easily it could be extracted. Then we propose a combined criterion, fl (frequent*length), to depict such synthesis effect on OOVs from both frequency and length.Figure 1 and 2 show the curves on OOVs between different values of frequent*length and the coverage from the CWS results generated by the enhanced model. At the starting part of the curves, we can see the OOVs are not well extracted. This is because the words in this part is either low frequent or short, or even with both conditions. When the words become longer or occurr more times, the coverage rapidly rises and goes steadily around 100%. It is well demonstrated that with a higher frequent*length value, the OOVs are more easily to be fully covered.

In spite of the better results, it is still necessary to consider the computational cost increasing by the additional unsupervised features. Taking it grant that the number of feature often decides the computational cost (time and space) in machine learning, the number of features for the training is counted.

Table 9. Comparisons of feature numbers

track	feature	feature number	increasing rate
CityU	N	8757120	-
	N+F	8758056	0.01%
UPUC	N	3815148	-
	N+F	3815994	0.02%

Table 9 shows the comparison on the feature numbers in different models, it is somewhat surprising that overall feature numbers of the enhanced model are just slightly more than that of the basic model. This fact suggests that although not many features are generated by unsupervised learning, they provide more discriminative information about the word boundary. Thus the results indicate that, the enhanced CRF model only costs the similar amount of computational resources as well as the basic n-gram CRF model. This result should be quite practical for any real application on CWS.

5. Conclusion

In this paper, we quantitatively analyze the effect of the unsupervised learning on character tagging approach for Chinese word segmentation. We measure the number of words generated by different models with and without enhancement from unsupervised learning on two corpora from SIGHAN Bakeoff-3. Both IVs and OOVs are analyzed by their frequency and length to find out what part are affected by the additional unsupervised features. We then investigate the reason why those features could help improve the supervised learning. And an empirical criterion, frequency*length is proposed to measure the quality of the OOVs which are extracted. It is seen that though unsupervised learning for supervised learning in CWS could benefit from those high frequent OOVs, it also accordingly helps improve the IVs. Moreover, no additional significant computational cost is required by the enhanced model in contrast to the basic model.

Acknowledgements

The research described in this paper was supported by City University of Hong Kong through the Strategic Research Grant 7002037 and 7002388.

References

[1] Nianwen Xue, Chinese word segmentation as character tagging, Computational Linguistics and Chinese Language Processing, 8(1):29C48. 2003.

[2] Hai Zhao, Chang-Ning Huang, and Mu Li, An improved Chinese word segmentation system with conditional random field. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pages 162C165, Sydney, Australia, July 22-23, 2006.

[3] Yan Song, Jiaqing Guo, Dongfeng Cai, Chinese Word Segmentation Based on an Approach of Maximum Entropy Modeling In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pages 162C165, Sydney, Australia, July 22-23, 2006.

[4] Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo, A Maximum Entropy Approach to Chinese Word Segmentation. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing (SIGHAN-4), pages 161C164, Jeju Island, Korea, October 14-15. 2005.

[5] John D. Lafferty, Andrew McCallum, and Fernando C. N, Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proceedings of the ICML2001, pages 282C289, San Francisco, CA.

[6] Gina-Anne Levow The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), Sydney, Australia, July, 2006

[7] Guangjin Jin and Xiao Chen The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6), Hyderabad, India, January 11-12, 2008

[8] Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. Accessor variety criteria for Chinese word extraction. Computational Linguistics, 30(1):75C93.

[9] Hai Zhao and Chunyu Kit, Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition, In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6), pp.106-111, Hyderabad, India, January 11-12, 2008

[10] Yan Song, Dongfeng Cai, Guiping Zhang, Hai Zhao, An Approach to Chinese Word Segmentation based on Character-Word Joint Decoding, Journal of Software (to appear), 2009

[11] Hai Zhao and Chunyu Kit, Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria for Chinese Word Segmentation, In Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008), Haifa, Israel, February 17-23, 2008.

[12] ChangNing Huang and Hai Zhao, Chinese Word Segmentation: A Decade Review, Journal of Chinese Information Processing, Vol. 21(3): 8-20, 2007

[13] Zhenxing Wang, Changning Huang and Jingbo Zhu, The Character-based CRF Segmenter of MSRA&NEU for the 4th Bakeoff, In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6), pp.106-111, Hyderabad, India, January 11-12, 2008

[14] Zhiyong Luo and Rou Song, An integrated method for Chinese unknown word extraction, In Proceedings of the Third SIGHAN Workshop on Chinese Language Processing, pages 148-154. Barcelona, Spain, 2004

[15] Aitao Chen, Chinese Word Segmentation Using Minimal Linguistic Knowledge, In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 148-151, 2003