Cost-benefit Analysis of Two-Stage Conditional Random Fields based English-to-Chinese Machine Transliteration

Chan-Hung Kuo[a] Shih-Hung Liu[ab] MikeTian-Jian Jiang[ac]
Cheng-Wei Lee[a] Wen-Lian Hsu[a]
[a]Institute of Information Science, Academia Sinica
[b]Department of Electrical Engineering, National Taiwan University
[c]Department of Computer Science, National Tsing Hua University
{laybow, journey, tmjiang, aska, hsu}@iis.sinica.edu.tw

Abstract

This work presents an English-to-Chinese (E2C) machine transliteration system based on two-stage conditional random fields (CRF) models with accessor variety (AV) as an additional feature to approximate local context of the source language. Experiment results show that two-stage CRF method outperforms the one-stage opponent since the former costs less to encode more features and finer grained labels than the latter.

1 Introduction

Machine transliteration is the phonetic transcription of names across languages and is essential in numerous natural language processing applications, such as machine translation, cross-language information retrieval/extraction, and automatic lexicon acquisition (Li et al., 2009). It can be either phoneme-based, grapheme-based, or a hybrid of the above. The phoneme-based approach transforms source and target names into comparable phonemes for an intuitive phonetic similarity measurement between two names (Knight and Graehl, 1998; Virga and Khudanpur, 2003). The grapheme-based approach, which treats transliteration as statistical machine translation problem under monotonic constraint, aims to obtain a direct orthographical mapping (DOM) to reduce possible errors introduced in multiple conversions (Li et al., 2004). The hybrid approach attempts to utilize both phoneme and grapheme information (Oh and Choi, 2006). Phoneme-based approaches are usually not good enough, because name entities have various etymological origins and transliterations are not always decided by pronunciations (Li et al., 2004). The state-of-the-art of transliteration approach is bilingual DOMs without intermediate phonetic projections (Yang et al., 2010).

Due to the success of CRF on sequential labeling problem (Lafferty et al., 2001), numerous machine transliteration systems applied it. Some of them treat transliteration as a two-stage sequential labeling problem: the first stage predicts syllable boundaries of source names, and the second stage uses those boundaries to get corresponding characters of target names (Yang et al., 2010; Qin and Chen, 2011). Dramatically de-creasing the cost of training with complex features is the major advantage of two-stage methods, but their downside is, compared to one-stage methods, features of target language are not directly applied in the first stage.

Richer context generally gains better results of sequential labeling, but squeezed performance always comes with a price of computational complexity. To balance cost and benefit for English-to-Chinese (E2C) transliteration, this work compares the one-stage method with the two-stage one, using additional features of AV (Feng et al., 2004) and M2M-aligner as an initial alignment (Jiampojamarn et al., 2007), to explore where the best investment reward is.

The remainder of this paper is organized as follows. Section 2 briefly introduces related works, including two-stage methods and AV. The machine transliteration system using M2M-aligner, CRF models, and AV features in this work is explained in Section 3. Section 4 describes

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 76–80,
Jeju, Republic of Korea, 8-14 July 2012.

experiment results along with a discussion in Section 5. Finally, Section 6 draws a conclusion.

2 Related Works

Reddy and Waxmonsky (2009) presented a phrase-based transliteration system that groups characters into substrings mapping onto target names, to demonstrate how a substring representation can be incorporated into CRF models with local context and phonemic information. Shishtla et al. (2009) adopted a statistical transliteration technique that consists of alignment models of GIZA++ (Och and Ney, 2003) and CRF models. Jiang et al. (2011) used M2M-aligner instead of GIZA++ and applied source grapheme's AV in a CRF-based transliteration.

A two-stage CRF-based transliteration was first designed to pipeline two independent processes (Yang et al., 2009). To recover from error propagations of the pipeline, a joint optimization of two-stage CRF method is then proposed to utilize n-best candidates of source name segmentations (Yang et al. 2010). Another approach to resist errors from the first stage is split training data into pools to lessen computation cost of sophisticated CRF models for the second stage (Qin and Chen, 2011).

3 System Description

3.1 EM for Initial Alignments

M2M-aligner first maximizes the probability of observed source-target pairs using EM algorithm and subsequently sets alignments via maximum *a posteriori* estimation. To obtain initial alignments as good as possible, this work empirically sets the parameter "maxX" of M2M-aligner for the maximum size of sub-alignments in the source side to 8, and sets the parameter "maxY" for the maximum size of sub-alignments in the target side to 1 (denoted as X8Y1 in short), since one of the well-known *a priori* of Chinese is that almost all Chinese characters are monosyllabic.

3.2 Format of Electronic Manuscript

The two-stage CRF method consists of syllable segmentation and Chinese character conversion CRF models, namely Stage-1 and Stage-2, respectively. Stage-1 CRF model is trained with source name segmentations initially aligned by M2M-aligner to predict syllable boundaries as accurate as possible. According to the discriminative power of CRF, some syllable boundary errors from preliminary alignments could be counterbalanced. Stage-2 CRF model then sees predicted syllable boundaries as input to produce optimal target names. For CRF modeling, this work uses Wapiti (Lavergne *et al.*, 2010).

Using "BULLOUGH" as an example, labeling schemes below are for Stage-1 training.

- B/*B* U/*B* L/*I* L/*I* O/*I* U/*I* G/*I* H/*E*
- B/*S* U/*B* L/*1* L/*2* O/*3* U/*4* G/*5* H/*E*

The first one is the common three-tag set "*BIE*". The last one is the eight-tag set "*B8*", including *B*, *1-5*, *E* and *S*: tag *B* indicates the beginning character of a syllable segment, tag *E* means the ending character, tag *I* or *1-5* stand for characters in-between, and tag *S* represents a single character segment. The expectation of the eight-tag set is the finer grained tags we used, the better segmentation accuracy we would gain.

For Stage-2, two labeling schemes are listed in the following.

- B/布 ULLOUGH/洛
- B/布 U/洛 L/*I* L/*I* O/*I* U/*I* G/*I* H/*I*

The former as substring-based labeling scheme are commonly used in two-stage CRF-based transliteration. Syllable segments in a source word are composed from Stage-1 results and then are associated with corresponding Chinese characters (Yang *et al.* 2009; Yang *et al.* 2010; Qin and Chen, 2011). The latter is a character-based labeling scheme where tags *B* or *S* from Stage-1 will be labeled with a Chinese character and others will be labeled as *I*. The merit of character-based method is to retrench the duration of the training, while substring-based method takes too much time to be included in this work for NEWS shared task. Section 5 will discuss more about pros and cons between substring and character based labeling schemes.

This work tests numerous CRF feature combinations, for example:

- C_{-3}, C_{-2}, C_{-1}, C_0, C_1 , C_2, C_3 and
- $C_{-3}C_{-2}$, $C_{-2}C_{-1}$, $C_{-1}C_0$, C_0C_1, C_1C_2, C_2C_3,

where local context is ranging from -3 to 3, and C_i denotes the characters bound individually to the prediction label at its current position *i*.

3.3 CRF with AV

AV was for unsupervised Chinese word segmentation (Feng *et al.*, 2004). Jiang *et al.*, (2011) showed that using AV of source grapheme as CRF features could improve transliteration. In our two-stage system, Source AV is used in Stage-1 in hope for better syllable segmentations, but not in Stage-2 since it may be redundant and surely increase training cost of Stage-2.

4 Experiment Results

4.1 Results of Standard Runs

Four standard runs are submitted to NEWS12 E2C shared task. Their configurations are listed in Table 1, where "U" and "B" denote observation combinations of unigram and bigram, respectively. A digit in front of a "UB", for example, "2", indicates local context ranging from -2 to 2. P_{BIE} stands for "BIE" tag set and P_{B8} is for "B8" tag set. To summarize, the 4th (*i.e.* the primary) standard run exceeds 0.3 in terms of top-1 accuracy (ACC), and other ACCs of standard runs are approximate to 0.3. The 3rd standard run uses the one-stage CRF method to compare with the two-stage CRF method. Experiment results show that the two-stage CRF method can excel the one-stage opponent, while AV and richer context also improve performance.

ID	Configuration	ACC	Mean F-score
1	Two-stage, 2UB, P_{BIE}	0.295	0.652
2	Two-stage, 2UB, P_{BIE}, AV	0.299	0.659
3	One-stage, 3UB, P_{BIE}, AV	0.291	0.654
4	Two-stage, 3UB, P_{B8}, AV	0.311	0.662

Table 1. Selected E2C standard runs

ID	Configuration	ACC	Mean F-score
I	Two-stage, 2UB, P_{BIE}, AV	0.363	0.707
II	Two-stage, 3UB, P_{B8}, AV	0.397	0.727
III	One-stage, 3UB, P_{BIE}, AV	0.558	0.834

Table 2. Selected E2C inside tests

ID	Number of Features	Numbers of Label
II	Stage-1: 60,496 Stage-2: 2,567,618	Stage-1: 8 Stage-2: 547
III	4,439,896	548

Table 3. Cost of selected E2C inside tests

4.2 Results of Inside Tests

Numerous pilot tests have been conducted by training with both the training and development sets, and then testing on the development set, as "inside" tests. Three of them are shown in Table 2, where configurations I and II use the two-stage method, and configuration III is in one-stage. Table 2 suggests a trend that the one-stage CRF method performs better than the two-stage one on inside tests, but Table 1 votes the opposite. Since the development set includes semi-semantic transliterations that are unseen in both the training and the test sets (Jiang *et al.*, 2011), models of inside tests are probably over-fitted to these noises. Table 3 further indicates that the number of features in the one-stage CRF method is doubled than that in the two-stage one. By putting these observations together, the two-stage CRF method is believed to be more effective and efficient than the one-stage CRF method.

5 Discussions

There are at least two major differences of two-stage CRF-based transliteration between our approach and others. One is that we enrich the local context as much as possible, such as using eight-tag set in Stage-1. The other is using a character-based labeling method instead of a substring-based one in Stage-2.

Reasonable alignments can cause CRF models troubles when a single source grapheme is mapped onto multiple phones. For instance, the alignment between "HAX" and "哈克斯" generating by M2M-aligner.

HA → 哈
X → 克斯

In this case, a single grapheme <X> pronounced as /ks/ in English therefore is associated with two Chinese characters "克斯", and won't be an easy case to common character-based linear-chain CRF. Although for the sake of efficiency, this work adopts character-based CRF models, only a few of such single grapheme for consonant blends or diphthongs appeared in training and test data, and then the decline of accuracy would be moderate. One may want to know how high the price is for using a substring-based method to solve this problem. We explore the number of features between substring-based and character-based

ID	Substring-based	Character-Based
II	106,070,874	2,567,618

Table 4. Number of features between substring and character based method in Stage-2

methods in Stage-2 with the same configuration II, as shown in Table 4. Features of substring-based method are tremendously more than character-based one. Qin (2011) also reported similar observations.

However, there is another issue in our character-based method: only the starting position of a source syllable segment will be labeled as Chinese character, others are labeled as *I*. Base on this labeling strategy, the local context of the target graphemes is missing.

6 Conclusions and Future Works

This work analyzes cost-benefit trade-offs between two-stage and one-stage CRF-based methods for E2C transliteration. Experiment results indicate that the two-stage method can outperform its one-stage opponent since the former costs less to encode more features and finer grained labels than the latter. Recommended future investigations would be encoding more features of target graphemes and utilizing *n*-best lattices from the outcome of Stage-1.

Acknowledgments

This research was supported in part by the National Science Council under grant NSC 100-2631-S-001-001, and the research center for Humanities and Social Sciences under grant IIS-50-23. The authors would like to thank anonymous reviewers for their constructive criticisms.

References

Haodi Feng, Kang Chen, Xiaotie Deng, and Wiemin Zheng. 2004. Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1):75-93.

Zellig Sabbetai Harris. 1970. Morpheme boundaries within words. Papers in Structural and Transformational Linguistics, 68-77.

Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. Proceedings of NAACL 2007, 372-379.

Mike Tian-Jian Jiang, Chan-Hung Kuo and Wen-Lian Hsu. 2011. English-to-Chinese Machine Transliteration using Accessor Variety Features of Source Graphemes. Proceedings of the 2011 Named Entities Workshop. 86-90.

K. Knight and J. Graehl. 1998. Machine Transliteration. Computational Linguistics, 24(4):599-612.

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of ICML, 591-598.

Thomas Lavergne, Oliver Cappé and François Yvon. 2010. Practical Very Large Scale CRF. Proceedings the 48th ACL, 504-513.

Haizhou Li, Min Zhang and Jian Su. 2004. A Joint Source Channel Model for Machine Transliteration. Proceedings of the 42nd ACL, 159-166.

Haizhou Li, A Kumaran, Min Zhang and Vladimir Pervouchine. 2009. Report of NEWS 2009 Transliteration Generation Shared Task. Proceedings of the 2009 Named Entities Workshop. 1-18.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19-51.

J. H. Oh and K. S. Choi. 2006. An Ensemble of Transliteration Models for Information Retrieval. Information Processing and Management, 42:980-1002.

Ying Qin. 2011. Phoneme strings based machine transliteration. Proceedings of the 7th IEEE International Conference on Natural Language Processing and Knowledge Engineering. 304-309.

Ying Qin and Guohua Chen. 2011. Forward-backward Machine Transliteration between English and Chinese Base on Combined CRF. Proceedings of the 2011 Named Entities Workshop. 82-85.

Eric Sven Ristad and Peter N. Yianilos. 1998. Learning String Edit Distance. IEEE Transactions on Pattern Recognition and Machine Intelligence, 20(5):522-532.

Sravana Reddy and Sonjia Waxmonsky. 2009. Substring-based transliteration with conditional random fields. Proceedings of the 2009 Named Entities Workshop, 92-95.

Praneeth Shishtla, V. Surya Ganesh, Sethuramalingam Subramaniam and Vasudeva Varma. 2009. A language-independent transliteration schema using character aligned models at NEWS 2009. Proceedings of the 2009 Named Entities Workshop, 40-43.

P. Virga and S. Khudanpur. 2003. Transliteration of Proper Names in Cross-lingual Information Retrieval. In the Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition.

Dong Yang, Paul Dixon, Yi-Cheng Pan, Tasuku Oonishi, Masanobu Nakamura, Sadaoki Furui. 2009. Combining a two-step conditional random field model and a joint source channel model for machine transliteration. Proceedings of the 2009 Named Entities Workshop, 72-75.

Dong Yang, Paul Dixon and Sadaoki Furui. 2010. Jointly optimizing a two-step conditional random field model for machine transliteration and its fast decoding algorithm. Proceedings of the ACL 2010. Conference Short Papers, 275-280

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.