Applying mpaligner to Machine Transliteration with Japanese-Specific Heuristics

Yoh Okuno
Job Hunter
nokuno@nokuno.jp

Abstract

We developed a machine transliteration system combining mpaligner (an improvement of m2m-aligner), DirecTL+, and some Japanese-specific heuristics for the purpose of NEWS 2012. Our results show that mpaligner is greatly better than m2m-aligner, and the Japanese-specific heuristics are effective for JnJk and EnJa tasks. While m2m-aligner is not good at long alignment, mpaligner performs well at longer alignment without any length limit. In JnJk and EnJa tasks, it is crucial to handle long alignment. An experimental result revealed that de-romanization, which is reverse operation of romanization, is crucial for JnJk task. In EnJa task, it is shown that mora is the best alignment unit for Japanese language.

1 Introduction

NEWS 2012 shared task regards transliteration as phonetic translation of proper nouns across different languages (Zhang et al., 2012). The most common approach for automatic transliteration is to follow the manner of statistical machine translation (Finch and Sumita, 2008). This approach mainly consists of 3 steps below.

1. Align training data monotonically
2. Train discriminative model given aligned data
3. Decode input characters to n-best candidate

One of the most popular alignment tools is m2m-aligner (Jiampojamarn et al., 2007), which is released as an open source software [1]. DirecTL+ (Jiampojamarn et al., 2008) is a decoding and training tool [2] and can be used with m2m-aligner for transliteration generation task.

However, m2m-aligner is not good at long alignment with no length limit. It tends to overfit for long alignment since its training is based on maximum likelihood estimation. Finch and Sumita (2010) proposed non-parametric Bayesian co-segmentation and applied it to machine transliteration (Finch et al., 2011). They penalized long alignment adopting Poisson distribution as prior of word length in the Bayesian model. Another method to penalize long alignment is proposed by Kubo et al. (2011) and released as mpaligner [3], originally developed for the purpose of Japanese pronunciation prediction. Just for its availability, we used mpaligner as an alternative of m2m-aligner.

Since m2m-aligner and mpaligner are both character-based alignment, there is a problem to produce phonetically invalid alignment. That is, character-based alignment may divide atomic units of characters, called mora, into meaningless pieces. Ideally, mora-to-mora alignment should be used for this task while no training data is provided for such purpose. In this paper, we propose Japanese-specific heuristics to cope with this problem depending on language-specific knowledge.

[1] http://code.google.com/p/m2m-aligner/
[2] http://code.google.com/p/directl-p/
[3] http://sourceforge.jp/projects/mpaligner/

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 61–65,
Jeju, Republic of Korea, 8-14 July 2012.

2 Related Works

Beside general researches for machine transliteration, there are other researches related to Japanese language. Cherry and Suzuki (2009) applied discriminative training to English-name-to-Japanese-Katakana transliteration. Hatori and Suzuki (2011) proposed a statistical machine translation approach for Japanese pronunciation prediction task. Hagiwara and Sekine (2011) used latent class model for transliteration including English-to-Japanese.

3 mpaligner: Minimum Pattern Aligner

mpaligner (Kubo et al., 2011) is an improvement of m2m-aligner. Their idea is simple; to penalize long alignment by scaling its probability using sum of their length. More formally, mpaligner uses a model;

$$P(x, y) = {p_{x,y}}^{|x|+|y|} \quad (1)$$

when deletion and insertion are not allowed. Here, x and y are source and target strings, $P(x, y)$ is probability of string pair (x, y), $p_{x,y}$ is a parameter which is estimated by previous iteration, and $|x|+|y|$ is sum of length of strings x and y. Though the scaled probability is no longer normalized, M-step of EM algorithm performs a kind of normalization.

4 Japanese-Specific Heuristics

Since mpaligner is a general-purpose alignment tool, we developed Japanese-specific heuristics as preprocessing for training data. That is, our system regards combined characters as one character, and applies mpaligner to them.

4.1 Romanized Japanese Name to Japanese Kanji Back-Transliteration Task (JnJk)

The most important heuristic for JnJk task is *deromanization*, which is the reverse operation of romanization. In Japanese language, consonants and vowels are coupled and expressed as Kana characters. Since Kana characters should not be divided, de-romanization converts romanized Japanese to Kana characters. This enables the system to align Kana character as minimal unit. For this conversion, a common romanization table for Japanese input method is used [4]. Moreover, a silent character called *Sokuon* is combined with its previous character since it can not be aligned alone.

Table 1 shows basic conversion table. We adopt longest-match algorithm to replace sequence of Roman characters to Kana characters. Without these operations, characters like "KA" may wrongly divided into "K" and "A" and aligned to different Kanji characters. More detailed examples are described in table 2. The bold rows are correct alignemnts performed by deromanization.

4.2 English to Japanese Katakana Task (EnJa)

In EnJa task, the alignment unit of target side should be mora, not character. For this purpose, our system combines lower case characters with their previous characters. Moreover, Japanese hyphen is also combined with the previous one since they form one mora.

As a result, "ァ", "ィ", "ゥ", "ェ", "ォ", "ャ", "ュ", "ョ", "ッ", "ー" are combined with their previous characters and treated as one mora. Table 3 shows alignment examples with and without this heuristics.

5 Experiments

In this section, we show the official scores for 8 language pairs and further investigation for JnJk and EnJa tasks.

5.1 Official Scores for 8 Language Pairs

Table 4 shows the official scores for 8 language pairs. In the official submits, we used mpaligner for alignment and DirecTL+ for training and decoding. We tried two version of mpaligner, 0.9 and 0.97, and chose better one as the primary submission. The version of DirecTL+ is 1.1, and the iteration number is selected automatically by the development set. For JnJk and EnJa tasks, we used our heuristics described above. For other language pairs, we just applied mpaligner and DirecTL+ using their default settings.

The results seem good, and we can find that ChEn, EnCh, EnHe and JnJk are difficult tasks in both measures ACC and F-Score.

[4] http://www.social-ime.com/romaji-table.html

Table 1: Basic De-romanization Table

Basic Romaji					
Roman	A	I	U	E	O
Kana	あ	い	う	え	お
Roman	KA	KI	KU	KE	KO
Kana	か	き	く	け	こ
Roman	SA	SI	SU	SE	SO
Kana	さ	し	す	せ	そ
Roman	TA	TI	TU	TE	TO
Kana	た	ち	つ	て	と
Roman	NA	NI	NU	NE	NO
Kana	な	に	ぬ	ね	の
Roman	HA	HI	HU	HE	HO
Kana	は	ひ	ふ	へ	ほ
Roman	MA	MI	MU	ME	MO
Kana	ま	み	む	め	も
Roman	YA		YU	YE	YO
Kana	や		ゆ	いぇ	よ
Roman	RA	RI	RU	RE	RO
Kana	ら	り	る	れ	ろ
Roman	WA	WI	WU	WE	WO
Kana	わ	うぃ	う	うぇ	を
Voiced Consonants (Dakuon)					
Roman	GA	GI	GU	GE	GO
Kana	が	ぎ	ぐ	げ	ご
Roman	ZA	ZI	ZU	ZE	ZO
Kana	ざ	じ	ず	ぜ	ぞ
Roman	DA	DI	DU	DE	DO
Kana	だ	ぢ	づ	で	ど
Roman	BA	BI	BU	BE	BO
Kana	ば	び	ぶ	べ	ぼ
Unvoiced Consonants (Han-Dakuon)					
Roman	PA	PI	PU	PE	PO
Kana	ぱ	ぴ	ぷ	ぺ	ぽ
Unvoiced Consonants (Yo-on)					
Roman	FA	FI	FU	FE	FO
Kana	ふぁ	ふぃ	ふ	ふぇ	ふぉ
Roman	SHA	SHI	SHU	SHE	SHO
Kana	しゃ	し	しゅ	しぇ	しょ
Roman	CHA	CHI	CHU	CHE	CHO
Kana	ちゃ	ち	ちゅ	ちぇ	ちょ

Table 2: Alignment Exapmles for JnJk Task

Unit	Source	Target
Roman	SUZ:UKI	鈴:木
Kana	**SUZU:KI**	**鈴:木**
Roman	HIR:OMI	裕:実
Kana	**HIRO:MI**	**裕:実**
Roman	OK:UNO	奥:野
Kana	**OKU:NO**	**奥:野**
Roman	JU:NYA	順:也
Kana	**JUN:YA**	**順:也**

Table 3: Alignment Exapmles for EnJa Task

Unit	Source	Target
Char	J:u:s:mi:ne	ジ:ャ:ス:ミ:ン
Mora	**Ju:s:mi:ne**	**ジャ:ス:ミ:ン**
Char	C:h:a:p:li:n	チ:ャ:ッ:プ:リ:ン
Mora	**Cha:p:li:n**	**チャッ:プ:リ:ン**
Char	A:r:th:ur	ア:ー:サ:ー
Mora	**Ar:thur**	**アー:サー**

Table 4: Official Scores for 8 Language Pairs

Task	ACC	F-Score	MRR	MAP
ChEn	0.013	0.259	0.017	0.013
EnBa	0.404	0.882	0.515	0.403
EnCh	0.301	0.655	0.376	0.292
EnHe	0.191	0.808	0.254	0.190
EnJa	0.362	0.803	0.469	0.359
EnKo	0.334	0.688	0.411	0.334
EnPe	0.658	0.941	0.761	0.640
JnJk	0.512	0.693	0.582	0.401

5.2 Investigation for JnJk Task

We further investigated the results for JnJk task to compare baseline and proposed system.

Table 5 shows the results of JnJk task for development set. The settings of tools are determined by preliminary experiments. We used m2m-aligner with length limit of maxX == 6 and maxY == 1, mpaligner with no length limit, and DirecTL+ with context size 7 and n-gram order 1. Proposed system is combined with Japanese-specific heuristics including de-romanization.

The results show two facts; mpaligner greatly beats m2m-aligner, and proposed de-romanization improves more both baseline systems.

Table 5: Results on JnJk Task

Method	ACC	F-Score	MRR	MAP
m2m-aligner	0.113	0.389	0.182	0.114
mpaligner	0.121	0.391	0.197	0.122
Proposed	**0.199**	**0.494**	**0.300**	**0.200**

5.3 Investigation for EnJa Task

In this subsection, we show the results for EnJa task to compare baseline and proposed system.

Table 6 shows the results of EnJa task for development set. All of the settings of tools are set default in this investigation.

Again, mpaligner beats m2m-aligner and our mora-based alignment improves scores of baseline systems in this system.

Table 6: Results on EnJa Task

Method	ACC	F-Score	MRR	MAP
m2m-aligner	0.280	0.737	0.359	0.280
mpaligner	0.326	0.761	0.431	0.326
Proposed	**0.358**	**0.774**	**0.469**	**0.358**

6 Disccussion

We compared mpaligner and m2m-aligner in the framework of statistical machine transliteration. In Japanese language, mpaligner performs better than m2m-aligner. This fact shows that maximum likelihood estimation approach adopted by m2m-aligner is not suitable for the purpose of machine transliteration. More importantly in practice, mpaligner is free from hand-tuning for length limits.

We proposed two Japanese-specific heuristics, de-romanization for JnJk task and mora-based alignment for EnJa task. They are implemented as preprocessing for training data, and improved the results of transliteration by eliminating linguistically invalid alignments. This shows the possibility that character-based alignment may not be the best solution for machine transliteration.

Beside Japanese, there can be efficient heuristics for other languages. But, more interesting issue is whether we can find such heuristics automatically or not.

7 Conclusion

We applied mpaligner to machine transliteration task for the first time and we proposed Japanese-specific heuristics for JnJk and EnJa tasks.

We confirmed that the maximum likelihood estimation approach adopted by m2m-aligner performs poor for the purpose of machine transliteration. One of methods to cope with this issue is to penalize long alignment using mpaligner.

We proposed de-romanization for JnJk task, and mora-based alignment for EnJa task. In the experiments, they demonstrated their capability to improve accuracy greatly.

Our proposed heuristics are language-dependent while they can be combined with any other language-independent methods including (Finch et al., 2011) or (Hagiwara and Sekine, 2011).

For future work, language-dependent heuristics beside Japanese or methods to find such heuristics automatically should be developed.

Acknowledgments

References

Colin Cherry and Hisami Suzuki. 2009. Discriminative substring decoding for transliteration. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1075, Singapore, August. Association for Computational Linguistics.

Andrew Finch and Eiichiro Sumita. 2008. Phrase-based machine transliteration. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific*

Speech Translation (TCAST), pages 13–18, Hyderabad, India, January.

Andrew Finch and Eiichiro Sumita. 2010. A Bayesian Model of Bilingual Segmentation for Transliteration. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 259–266.

Andrew Finch, Paul Dixon, and Eiichiro Sumita. 2011. Integrating models derived from non-parametric bayesian co-segmentation into a statistical machine transliteration system. In *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, pages 23–27, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Masato Hagiwara and Satoshi Sekine. 2011. Latent class transliteration based on source language origin. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 53–57, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jun Hatori and Hisami Suzuki. 2011. Japanese pronunciation prediction as phrasal statistical machine translation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 120–128, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 905–913, Columbus, Ohio, June. Association for Computational Linguistics.

Keigo Kubo, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano. 2011. Unconstrained many-to-many alignment for automatic pronunciation annotation. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2011 (APSIPA2011)*, Xi'an, China, October.

Min Zhang, A Kumaran, and Haizhou Li. 2012. Whitepaper of news 2012 shared task on machine transliteration. In *Proceedings of the 4th Named Entities Workshop (NEWS 2012)*, Jeju, Korea, July. The Association of Computational Linguistics.