Language Independent Transliteration system using phrase based SMT approach on substrings

Sara Noeman

IBM Cairo Technology & Development Center
Giza, Egypt
noemans@eg.ibm.com

Abstract

Everyday the newswire introduce events from all over the world, highlighting new names of persons, locations and organizations with different origins. These names appear as *Out of Vocabulary* (OOV) words for Machine translation, cross lingual information retrieval, and many other NLP applications. One way to deal with OOV words is to *transliterate* the unknown words, that is, to render them in the orthography of the second language.

We introduce a statistical approach for transliteration only using the bilingual resources released in the shared task and without any previous knowledge of the target languages. Mapping the Transliteration problem to the Machine Translation problem, we make use of the *phrase based SMT* approach and apply it on *substrings* of names. In the English to Russian task, we report ACC (*Accuracy in top-1*) of 0.545, Mean F-score of 0.917, and MRR (*Mean Reciprocal Rank*) of 0.596.

Due to time constraints, we made a single experiment in the English to Chinese task, reporting ACC, Mean F-score, and MRR of 0.411, 0.737, and 0.464 respectively.

Finally, it is worth mentioning that the system is language independent since the author is not aware of either languages used in the experiments.

1. Introduction

Named entities translation is strongly required in the field of Information retrieval (IR) as well as its usage in Machine translation. A significant proportion of OOV words are named entities and typical analyses find around 50% of OOV words to be named entities, yet these can be the most important words in the queries. Larkey et al (2003) showed that average precision of cross language retrieval reduced more than 50% when named entities in the queries were not translated.

Transliteration may be considered as a phonetic translation or mapping of a *sequence of characters* in the source language in the alphabet of the target language, thus we can use the analogy with the Machine translation problem, which translates a *sequence of words* in the source language into a semantically equivalent *sequence of words* in the target language.

In a statistical approach to machine translation, given a foreign word F, we try to find the English word Ê that maximizes P(E\F). Using Bayes' rule, we can formulate the task as follows:

$$\hat{E} = \underset{E}{\operatorname{argmax}} \frac{P(F\backslash E)*P(E)}{P(F)}$$

$$= \underset{E}{\operatorname{argmax}}\ P(F\backslash E)*P(E)$$

This is known as the noisy channel model, which splits the problem into two sub-tasks. The translation model provides an estimate for the P(F\E) for the foreign word ***F*** being a translation for the English word ***E***, while the language model provides an estimate of the probability P(E) is an English word.

In this paper we use the phrase based statistical Machine Translation (PBSMT) approach introduced by (Koehn et al.) to build English to Russian, and English to Chinese transliteration systems capable of learning the substring to substring mapping between source and target languages.

Section 2 includes a detailed description of our approach, section 3 describes our experimental set up and the results. The conclusions and future work are explained in section 4.

2. System architecture

Our approach is a formulation of the Transliteration problem using the PBSMT technique that proved improvement in Machine translation domain, making use of the analogy between the two problems.

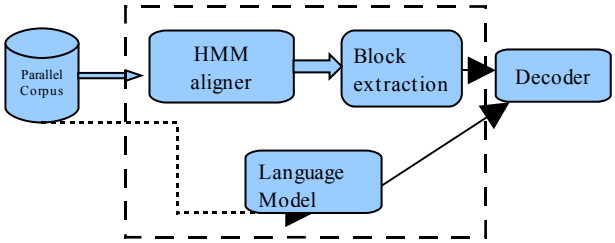
The phrase-based approach developed for statistical machine translation (Koehn et al., 2003) is designed to overcome the restrictions of many-to-many mappings in word-based translation models. We applied the phrase based statistical approach used in Machine translation on our problem, mapping the "word", and

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 112–115,
Suntec, Singapore, 7 August 2009.

"phrase" in PBSMT terminology into "character", and "substring" in our system, where the substring in our notation represents a *sequence of adjacent characters*.

Figure (1) shows an overview of the whole system architecture.

Figure (1)



We used an HMM aligner similar to Giza++ (Och. et al., 1999) over the parallel character sequences using forward-backward alignment intersection. Heuristics were used to extend substring to substring mappings based on character-to-character alignment, with the constraint that no characters within the substring pair are linked to characters outside the substring pair. Thus we generated a substring to substring translation model with relative frequencies. We deploy heuristics to extract character sequence mapping similar to the heuristics used in PBSMT (Koehn et al., 2003). Figure (2) shows the heuristics used for block extraction over substrings in the English to Russian task using character to character alignments.

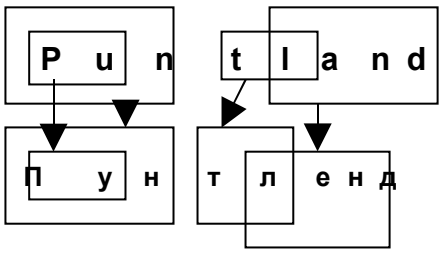


Figure (2)

Unlike the Machine Translation task, in transliteration we do not need any reordering during decoding which makes the decoding phase easier. We used monotone beam search decoder generating the *best k* transliteration candidates, where the translation model and the language model are used by the decoder to get best Viterbi paths of character sequences as a phonetic translation for the input English character sequence. (Tillmann, et al., 2003).

Finally, all transliteration candidates are weighted using their translation and language model probabilities as follows:

$P(w_r \setminus w_e) = P(w_e \setminus w_r) * P(w_r \in R)$

Here, we explain our system for the English to Russian task, while the English to Chinese system will follow the same criteria and their results are mentioned later.

a. Data and Resources

Standard Runs:

In the English to Russian task, we used the parallel corpus (EnRu) released by NEWS 2009 Shared Task on Transliteration to build the translation model. For the English to Chinese standard run, we used the parallel English-Chinese (EnCh) corpus released by NEW2009 availed by (Li et al., 2004). The target language side (Russian, Chinese) of the parallel data was used to build the language model. NEWS2009 released 5977 of EnRu names pairs as a training set, and 943 pairs as a development set. The EnCh corpus had 31,961 pairs as a training set, and 2896 pairs as a development set.

Non-Standard Runs:

For the English to Russian task we used the Russian data in UMC 0.1 Czech-English-Russian, from the Institute of Formal and Applied Linguistics (ÚFAL), to build a larger Russian LM, in addition to the data resources used in the standard run. No Named Entity tagging has been applied on this data because we lack the tools. However, we are just validating the character n-gram sequences in the target language with larger corpus of character sequences.

We didn't use any additional resources for the Chinese task.

b. Training

The training is held in two phases; first learning the list of Russian characters aligned to multiple English characters, and thus we obtain a table of English character n-grams to be added to unigram inventory of the source language. The second stage learns the transliteration model over this new inventory. (Larkey et al., 2003).

Table 1 shows the list of English n-gram characters added to unigram inventory.

Table (1)

s h c h	shch
s z c z	szcz
s c h	sch
z h	zh
c k	ck
p h	ph
k h	kh
c h	ch
s h	sh
s z	sz
c z	cz
š č	šč

A substring (phrase) table of Russian substrings mapped to English substrings is considered as the

translation model P(E\R). A language model P(R) is built using a monolingual Russian corpus.
Figure (3) shows a sample of the *substring feature table* generated during training using the block extraction heuristics over HMM alignments.

а к о н ‖ e a c o n 0 1
а ф ‖ e a f 0 1
а ф э ‖ e a fă 0 1
е н е р и ф ‖ e n e r i f 0 1
е н е р и ф е ‖ e n e r i f e 0 1
н е р с ‖ e n e r s 0 1
н е р с р ‖ e n e r s r 0 1
н е р с р ю ‖ e n e r s rü 0 1
Figure (3) a sample of the substring table

c. Decoding

The source English word is fragmented into all its possible substring sequences, and the decoder applies a monotone beam search, without reordering, to generate the *best k* phonetic translation character sequences in the target language alphabet.
Experiments 1, 2, and 3 use a substring based transliteration system The experiments set up will be as follows:

i. The effect of true casing versus lowercasing Russian characters is explained through the first experiment (*Exp-1*).

ii. The released English data contains some unusual English characters not belonging to the English alphabet, some of which are vowels like "ế, ê, ē, é, ä, ã, ą, ö, ó, ō, ú, û, ŭ", and others are consonants as "ľ, ł, ť, ż, ž, ń, ñ, ň, ř". The effect of normalizing these unusual English characters is explained in the second experiment (*Exp-2*).

iii. In the third experiment (*Exp-3*) we used the unigram inventory described in Table (1).

N.B.: Chinese language has a very large number of characters representing syllables rather than characters (a syllables = a consonant + vowel, or a consonant + vowel + final), thus the unigram inventory used in the English to Chinese task wasn't generated using the statistical trend used with English-Russian task. General linguistic heuristics were used to re-merge character n-grams like "sh, th, gh, ph, etc…" as well as character repetitions like "ll, mm, nn … ss, tt, etc..."

3. Results

Evaluation Metrics:
The quality of the transliteration task was measured using the 6 metrics defined in the shared task white paper. The first metric is the *Word Accuracy in Top-1 (**ACC**)* which is the precision of the exact match with the Top-1 reference. The second one is the *Fuzziness in Top-1 (**Mean F-score**)* which reflects an average F-score of the normalized lowest common subsequence between the system output and the Top-1 reference. The ***(MRR)*** represents the *Mean Reciprocal Rank* of the Top-1 reference in the ***k*** candidates generated by the system. The last three metrics MAP_{ref}, MAP_{10}, MAP_{sys} measure how the ***k*** candidates generated by the transliteration system are mapped to the ***n*** references available for each input in the testset.

English to Russian task
The results of experiments 1, 2, and 3 on the Development set, using the 6 evaluation metrics explained before, are written in Table (2). Exp-2 reflects the effect of normalizing all the unusual English characters that existed in the training data. Referring to the results of Exp-1, we conclude that this normalization decreases the ACC of the system around 2.5%. In the next experiments we only use the set up of Exp-3, which uses the statistical unigram inventory without true casing Russian characters or normalizing unusual English characters.

	Exp-1	Exp-2	Exp-3
ACC	0.705	0	0
Mean F-score	0.945	0.939	0
MRR	0.741	0.721	0
MAPref	0.705	0	0
MAP10	0.220	0.215	0
MAPsys	0.525	0	0

Table (2) explains Eng-Russian task results on the Development Set for experiments 1, 2, and 3.

- **Standard Run:**

Our Standard Run submission used the same setup used in Experiment-3, no lowercasing, no normalization, and using the list of English n-grams that were added to the unigram inventory after the first training phase. Table (3) contains the results of our Standard Submissions.

	Standard submission
ACC	0.545
Mean F-score	0.917
MRR	0.596
MAP_{ref}	0.545
MAP_{10}	0.286
MAP_{sys}	0.299

Table (3) explains Eng-Russian task results on the blind Test Set. This was the Standard submission.
N.B.: We submitted the previous output in true-cased Russian characters as our standard submission, and then we submitted the same system output after lower casing as a Non-Standard run because we were not sure that the evaluation tool used by the Shared Task

will be able to map true case and lower case variations.
The same will be done in the next run, where 2 submissions are submitted for the same output, one of which was true-cased and the other was lower cased.

- **Non-Standard Run:**

Using (*UMC 0.1*) additional LM on the blind Test set. The results are in table(5)

	Non-Standard submission
ACC	0.524
Mean F-score	0.913
MRR	0.579
MAP_{ref}	0.524
MAP_{10}	0.277
MAP_{sys}	0.291

Table (5) explains Eng-Russian task results on the blind Test Set. This was the Non-Standard submission.

English to Chinese task

Finally the previous setup with slight modifications was applied to the Eng-Chinese transliteration task. Tables (6), and (7) represent the results on the Chinese Development set and Test set respectively.

	Exp-3
ACC	0.447
Mean F-score	0.748
MRR	0.489
MAP_{ref}	0.447
MAP_{10}	0.147
MAP_{sys}	0.191

Table (6) explains Eng-Chinese task results on the Development Set.

- **Standard Run:**

	Standard submission
ACC	0.411
Mean F-score	0.737
MRR	0.464
MAP_{ref}	0.411
MAP_{10}	0.141
MAP_{sys}	0.173

Table (7) explains Eng-Chinese task results on the blind Test Set. This was the Standard submission

4. Conclusion and Future Work

In this paper we presented a substring based transliteration system, making use of the analogy between the Machine translation task and Transliteration. By applying the phrase based SMT approach in the transliteration domain, and without any previous knowledge of the target languages, we built an English to Russian system with ACC of 54.5% and an English to Chinese system with ACC of 41.2%.

In the future we are planning to hold some experiments to filter out the generated phrase table (substring table) and try other decoding techniques.

5. Acknowledgement

I would like to thank Dr. Hany Hassan in IBM Cairo TDC for his helpful comments and technical support.

6. References

N. AbdulJaleel and L. S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In *CIKM*, pages 139–146.

Y. Al-Onaizan and K. Knight. 2002. Machine Transliteration of Names in Arabic Text. In Proceed-ings of the ACL Workshop on Computational Approaches to Semitic Languages.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263–311.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. *Proc. Of the Human Language Technology Conference*, *HLT-NAACL'2003*, May.

H. Li, M. Zhang, J. Su: A Joint Source-Channel Model for Machine Transliteration. ACL 2004: 159-166

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In June 1999, EMNLP.

T. Sherif and G. Kondrak. 2007. Substring-Based Transliteration. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.

C. Tillmann and H. Ney. 2003.Word Re-ordering and DP-based Search in Statistical Machine Translation. In *COLING*, pages 850-856.

J. Zobel and P. Dart. 1996. Phonetic String Matching. Lessons from Information Retrieval. SIGIR Forum, special issue:166—172.