Machine Transliteration

Mohamed Abdel Fattah [1], Fuji Ren [1, 2], Shingo Kuroiwa [1]
[1]. Faculty of Engineering, University of Tokushima
2-1 Minamijosanjima
Tokushima, Japan 770-8506
[2]. School of Information Engineering, Beijing University of Posts & Telecommunications
Beijing, 100088, China
(mohafi, ren, kuroiwa) @is.tokushima-u.ac.jp

Abstract. In the present study, we present different approaches for transliteration proper noun pair's extraction from parallel corpora based on different similarity measures between the English and Romanized Arabic proper nouns under consideration. The strength of our new system is that it works well for low-frequency words. We evaluate the presented new approaches using an English-Arabic parallel corpus. Most of our results outperform previously published results in terms of precision, recall and F- Measure.

1 Introduction

Recently, much research has been done on machine transliteration for many language pairs, such as English/Arabic [1] and English/Korean [2]. Most of the above approaches require a pronunciation dictionary for converting a source word into a sequence of pronunciations. However, words with unknown pronunciations may cause problems for transliteration. On the other hand, much research has focused on the study of automatic bilingual lexicon construction based on bilingual corpora. Proper names and corresponding transliterations can often be found in parallel corpora or topic-related bilingual comparable corpora. However, many methods dealt with this problem based on the frequencies of words appearing in corpora, an approach which cannot be effectively applied to low-frequency words. Fung, used different approaches to create translation pairs from parallel and comparable corpora [3]. We have exploited the pattern matching method of [3] to extract transliteration pairs from English – Arabic parallel corpus and we used it as a base line method.

2 The Arabic-English Parallel Corpus

We have applied our transliteration techniques on the "Arabic English Parallel News Text Part 1", Linguistic Data Consortium (LDC) catalog number LDC2004T18 and ISBN 1-58563-310-0. [4].

3 Patten Matching Approach

We treat the transliteration compilation problem as a pattern matching problem in [3] with little modification in the first step to decrease the computation time. In the first step of the algorithm, we did not tag the English half of the parallel text only but we also tagged the Arabic half in order to restrict the matching process on as few words as possible to decrease the computation time. We achieved accuracy and recall of 71.4% 66.5% respectively for the best matched pairs. We also achieved accuracy and recall of 73.8% and 68.2% respectively for the top three Arabic transliterations for an English proper noun respectively. We found that many mistaken transliterations resulted from insufficient data.

4. The Proposed English-Arabic Proper Noun Transliteration Pairs Creation Approach

The system extracts all proper nouns from the English sentence using the CLAWS4 POS tagger. It also extracts all proper nouns from the associated Arabic sentence using the Buckwalter Arabic Morphological Analyzer Version 1.0. All the Arabic proper nouns are romanized using [5]. The similarity between every English and Romanized Arabic proper noun pair is measured. The English-Arabic proper noun pair which has similarity score above certain threshold (th) is extracted. The system repeats this step for all English and Arabic proper nouns exist in the sentence pair. The system applies the previous steps on all remaining sentence pairs.

4.1. Experimental Results using Dice's Similarity Coefficient

Apply the new approach on the English- Arabic corpus, and use Dice's Similarity Coefficient to measure the similarity between the English proper noun and the Romanized Arabic proper noun. Table 1 shows the precision, recall and the harmonic mean of precision and recall (F-Measure) for the transliteration pairs extracted as a function of the threshold "th".

4.2. Experimental Results using SIM1

Most of Arabic words have a syllable of CV. Most of the Arabic words contain short or long vowel between two consonant letters. Take the Arabic word "محمد" "mohammad" as an example. The short vowels 'o', 'a' and 'a' are existed between the consonants "m, h", "h, m" and "m, d" respectively. Moreover the short vowels are not appeared on the Arabic words in almost all Arabic documents. Hence, the Dice's approach to measure similarity between English- Arabic transliteration pairs does not work well. We have decided to use our proposed similarity measure called "SIM1". We use the following algorithm to specify "SIM1":

```
  Set SIM1 = 0
  Set ia = ie = 0
  R:   Read the Romanized Arabic character(ia)
      Read the English character(ie)
      If (the Romanized Arabic character(ia) = the English
character(ie))
            SIM1 = SIM1 + 1
      End
      Else ie = ie + 1 & Read the English character(ie)
            If (the Romanized Arabic character(ia) = the English
      character(ie))
                  SIM1 = SIM1 + 1
            End
            Else If(English character(ie - 1)= English character(ie)))
            ie = ie + 1 & Read the English character(ie)
        If (the Romanized Arabic character(ia) = the English
        character(ie))
                              SIM1 = SIM1 + 1
                        End
              End
      End
```

Table 1: the results using Dice's Similarity Coefficient

th	1.0	0.9	0.8	0.7	0.6	0.5	0.3	0.0
Precision	100%	100%	95.9%	86.7%	72.1%	61.3%	42.8%	24.2%
Recall	2.3%	2.3%	6.2%	15.3%	22.6%	28.7%	36.5%	98.1%
F-Measure	4.5%	4.5%	11.6%	26.0%	34.4%	39.1%	39.4%	38.8%

```
    ia = ia + 1 & ie = ie + 1
    if (ia < Length (Romanized Arabic word))
          GOTO R
    End
SIM1 = SIM1/(max_Length(Romanized Arabic word, English word))
```

Table 2 shows the results when we apply the previous algorithm to specify SIM1 as a similarity score for the transliteration pair under consideration. It is clear from table 2 that the recall has been improved compared with table 1.

4.3. Experimental Results using SIM2

As we notice in the previous section, in the transliteration pair "aladl, الامل", when the Arabic word "الامل" is converted to English alphabet, it will be "alaml". If we match "aladl" with "alaml", only 'd' and 'm' do not match. So the similarity score SIM1 = 0.8. And the pair is not correct. Hence, it is required that the system restricts the extracted transliteration pairs only on the pairs that have all Romanized Arabic characters matched with some or all English proper noun characters to increase the precision. We achieve that by modifying the previous algorithm to set the similarity score to zero if any Romanized Arabic character does not match with any English character. Hence we use a new similarity measure called SIM2. Using SIM2 as a similarity measure, we achieved the results in table 3.

5 Conclusions and Future Work

In this study, we presented a new system to create English- Arabic transliteration pairs from parallel corpora based on different similarity measure approaches. The strength of our new system is that it works well for low-frequency words. The system could extract some correct transliteration pairs of frequency equal to 1. We found that the similarity measure must be specified based on the characteristics of the two languages pair under consideration. We have evaluated the presented new approaches using an English- Arabic parallel corpus.

Table 2: the results using SIM1 Similarity Coefficient

th	1.0	0.9	0.8	0.7	0.6	0.5	0.3	0.0
Precision	100%	100%	92.4%	75.7%	61.8%	36.3%	22.1%	24.2%
Recall	2.3%	6.5%	26.7%	45.2%	57.1%	62.4%	78.7%	98.1%
F-Measure	4.5%	12.2%	41.4%	56.6%	59.4%	45.9%	34.5%	38.8%

Table 3: the results using SIM2 Similarity Coefficient

th	1.0	0.9	0.8	0.7	0.6	0.5	0.3	0.0
Precision	100%	100%	98.9%	94.5%	88.6%	56.1%	34.2%	24.2%
Recall	2.3%	6.5%	24.6%	42.3%	53.4%	60.3%	66.4%	98.1%
F-Measure	4.5%	12.2%	39.4%	58.4%	66.6%	58.1%	45.1%	38.8%

In a future work, we will use the resulted transliteration pairs in cross language information retrieval and machine translation systems.

Acknowledgment: This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14380166 and 17300065, Exploratory Research 17656128 in 2005, International Communications Foundation (ICF).

References

1. Al-Onaizan, Y., Knight, K.: Translating named entities using monolingual and bilingual resources. ACL, Philadelphia, (2002) 400–408.
2. Kang, B. J., Choi, K. S.: Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. International Journal of Computer Processing of Oriental Languages 14 (2) (2001) 109–131.
3. Fung, P.: A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. ACL, (1995) 236-243.
4. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T18
5. http://archimedes.fas.harvard.edu/mdh/arabic/arabic-loc.pdf