An Integrated Approach for Arabic-English Named Entity Translation

Hany Hassan
IBM Cairo Technology Development Center
Giza - Egypt
P.O. Box 166 Al-Ahram
hanyh@eg.ibm.com

Jeffrey Sorensen
IBM T.J. Watson Research Center
Yorktown Heights
NY 10598
sorenj@us.ibm.com

Abstract

Translation of named entities (NEs), such as person names, organization names and location names is crucial for cross lingual information retrieval, machine translation, and many other natural language processing applications. Newly named entities are introduced on daily basis in newswire and this greatly complicates the translation task. Also, while some names can be translated, others must be transliterated, and, still, others are mixed. In this paper we introduce an integrated approach for named entity translation deploying phrase-based translation, word-based translation, and transliteration modules into a single framework. While Arabic based, the approach introduced here is a unified approach that can be applied to NE translation for any language pair.

1 Introduction

Named Entities (NEs) translation is crucial for effective cross-language information retrieval (CLIR) and for Machine Translation. There are many types of NE phrases, such as: person names, organization names, location names, temporal expressions, and names of events. In this paper we only focus on three categories of NEs: person names, location names and organization names, though the approach is, in principle, general enough to accommodate any entity type.

NE *identification* has been an area of significant research interest for the last few years. NE translation, however, remains a largely unstudied problem. NEs might be phonetically transliterated (e.g. persons names) and might also be mixed between phonetic transliteration and semantic translation as the case with locations and organizations names.

There are three distinct approaches that can be applied for NE translation, namely: a transliteration approach, a word based translation approach and a phrase based translation approach. The transliteration approach depends on phonetic transliteration and is only appropriate for out of vocabulary and completely unknown words. For more frequently used words, transliteration does not provide sophisticated results. A word based approach depends upon traditional statistical machine translation techniques such as IBM Model1 (Brown et al., 1993) and may not always yield satisfactory results due to its inability to handle difficult many-to-many phrase translations. A phrase based approach could provide a good translation for frequently used NE phrases though it is inefficient for less frequent words. Each of the approaches has its advantages and disadvantages.

In this paper we introduce an integrated approach for combining phrase based NE translation, word based NE translation, and NE transliteration in a single framework. Our approach attempts to harness the advantages of the three approaches while avoiding their pitfalls. We also introduce and evaluate a new approach for aligning NEs across parallel corpora, a process for automatically extracting new NEs translation phrases, and a new transliteration approach. As is typical for statistical MT, the system requires the availability of general parallel corpus and Named Entity identifiers for the NEs of interest.

Our primary focus in this paper is on translating NEs out of context (i.e. NEs are extracted and translated without any contextual clues). Although

Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 87–93,
Ann Arbor, June 2005.

this is a more difficult problem than translating NEs in context, we adopt this approach because it is more generally useful for CLIR applications.

The paper is organized as follows, section 2 presents related work, section 3 describes our integrated NE translation approach, section 4 presents the word based translation module, the phrase based module, the transliteration module, and system integration and decoding, section 5 provides the experimental setup and results and finally section 6 concludes the paper.

2 Related Work

The Named Entity translation problem was previously addressed using two different approaches: Named Entity phrase translation (which includes word-based translation) and Named Entity transliteration. Recently, many NE phrase translation approaches have been proposed. Huang et al. (Huang et al., 2003) proposed an approach to extract NE trans-lingual equivalences based on the minimization of a linearly combined multi-feature cost. However this approach used a bilingual dictionary to extract NE pairs and deployed it iteratively to extract more NEs. Moore (Moore, 2003), proposed an approach deploying a sequence of cost models. However this approach relies on orthographic clues, such as strings repeated in the source and target languages and capitalization, which are only suitable for language pairs with similar scripts and/or orthographic conventions.

Most prior work in Arabic-related transliteration has been developed for the purpose of machine translation and for Arabic-English transliteration in particular. Arbabi (Arbabi et al., 1998) developed a hybrid neural network and knowledge-based system to generate multiple English spellings for Arabic person names. Stalls and Knight (Stalls and Knight, 1998) introduced an approach for Arabic-English back transliteration for names of English origin; this approach could only back transliterate to English the names that have an available pronunciation. Al-Onaizan and Knight (Al-Onaizan and Knight, 2002) proposed a spelling-based model which directly maps English letter sequences into Arabic letter sequences. Their model was trained on a small English Arabic names list without the need for English pronunciations. Although this method does not require the availability of English pronunciation, it has a serious limitation because it does not provide a mechanism for inserting the omitted short vowels in Arabic names. Therefore it does not perform well with names of Arabic origin in which short vowels tend to be omitted.

3 Integrated Approach for Named Entity Translation

We introduce an integrated approach for Named Entity (NE) translation using phrase based translation, word based translation and transliteration approaches in a single framework. Our unified approach could handle, in principle, any NE type for any languages pair.

The level of complication in NE translation depends on the NE type, the original source of the names, the standard *de facto* translation for certain named entities and the presence of acronyms. For example persons names tend to be phonetically transliterated, but different sources might use different transliteration styles depending on the original source of the names and the idiomatic translation that has been established. Consider the following two names:

“جاك شيراك : jAk $yrAk” → “Jacques Chirac”
“جاك سترو :jAk strw” → “Jack Straw”

Although the first names in both examples are the same in Arabic, their transliterations should be different. One might be able to distinguish between the two by looking at the last names. This example illustrates why transliteration may not be good for frequently used named entities. Transliteration is more appropriate for unknown NEs.

For locations and organizations, the translation can be a mixture of translation and transliteration. For example:

شركة مايكروسوفت :$rkp mAykrwswft → Microsoft Company
القدس : Alqds → Jerusalem
مطار طوكيو: mTAr Tokyw → Tokyo Airport

These examples highlight some of the complications of NE translation that are difficult to overcome using any phrase based, word based or transliteration approach independently. An approach that integrates phrase and word based translation with transliteration in a systematic and flexible framework could provide a more complete solution to the problem.

Our system utilizes a parallel corpus to separately acquire the phrases for the phrase based system, the translation matrix for the word based system, and training data for the transliteration system. More details about the three systems will be presented in the next section. Initially, the corpus is automatically annotated with NE types in the source and target languages using NE identifiers similar to the systems described in (Florian et al., 2004) for NE detection.

4 Translation and Transliteration Modules

4.1 Word Based NE Translation

- **Basic multi-cost NE Alignment**

We introduce a novel NE alignment technique to align NEs from a parallel corpus that has been automatically annotated with NE types for source and target languages. We use IBM Model1, as introduced in (Brown et. al, 1993), with a modified alignment cost. The cost function has some similarity with the multi-cost aligning approach introduced by Huang (Huang et al. 2003) but it is significantly different. The cost for aligning any source and target NE word is defined as:

$$C = \lambda_1 p(w_e | w_f) + \lambda_2 Ed(w_e, w_f) + \lambda_3 Tag(w_e, w_f)$$

Where: w_e and w_f are the target and source words respectively and λ_1, λ_2 and λ_3 are the cost weighting parameters.

The first term $p(w_e | w_f)$ represents the translation log probability of target word (w_e) given the source word (w_f). The second term $Ed(w_e, w_f)$ is length-normalized phonetic based edit distance between the two words. This phonetic-based edit distance employs an Editex style (Zobel and Dart, 1996) distance measure, which groups letters that can result in similar pronunciations, but doesn't require that the groups be disjoint, and can thus reflect the correspondences between letters with similar pronunciation more accurately. The Editex distance (*d*) between two letters *a* and *b* is:

d(*a*,*b*) = 0 if both are identical
= 1 if they are in the same group
= 2 otherwise

The Editex distance between two words is the summation of Editex distance between their letters and length-normalized edit distance is:

$$Ed(w_e, w_f) = \log(1 - \frac{d(w_e, w_f)}{\max(|w_e|, |w_f|)})$$

where $d(w_e, w_f)$ is the "Editex" style edit distance and $\max(|w_e|, |w_f|)$ is the maximum of the two lengths for the source and target, normalizing the edit distance.

The Editex edit distance is deployed between English words and "romanized" Arabic words with a grouping of similar consonants and a grouping of similar vowels. This helps in identifying the correspondence between rare NEs during the alignment. For example, consider two rare NE phrases that occur once in the training:

"*وقد استدعى وزير الخارجية الياباني نوبوتاكا ماشيمورا السفير الصيني وانج يي*"

"*wqd AstdEY wzyr AlxArjyp AlyAbAny nwbwtAkA mA$ymwrA Alsfyr AlSyny wAnj yy*"

"*Japanese Foreign Minister Nobutaka Machimura has summoned the Chinese ambassador Wang Yee*"

Thus the task of the alignment technique is to align
نوبوتاكا :nwbwkAtA → Nobutaka
ماشيمورا : mA$ymwrA → Machimura
وانج :wAng → Wang
يي :yy → Yee

If a pure Model-1 alignment was used, then the model would have concluded that all words could be aligned to all others with equal probability. However, the multi-cost alignment technique could align two named entities using a single training sample. This approach has significant effect in correctly aligning rare NEs.

The term $Tag(w_e, w_f)$ in the alignment cost function is the NE type cost which increases the alignment cost when the source and target words are annotated with different types and is zero otherwise.

The parameters of the cost function ($\lambda_1, \lambda_2, \lambda_3$) can be tuned according to the NE category and to frequency of a NE. For example, in the case of person's names, it might be advantageous to use a larger λ_2 (boosting the weight of transliteration).

- **Multi-cost Named Entity Alignment by Content Words Elimination**

In the case of organization and location names; many content words, which are words other than the NEs, occur in the NE phrases. These content words might be aligned incorrectly to rare NE words. A two-phase alignment approach is deployed to overcome this problem. The first phase is aligning the content words using a content-word-only translation matrix. The successfully aligned content words are removed from both the source and target sentences. In the second phase, the remaining words are subsequently aligned using the multi-cost alignment technique described in the previous section. This two-phase approach filters out the words that might be incorrectly aligned using the single phase alignment techniques. Thus the alignment accuracy is enhanced; especially for organization names since organization names used to contain many content words.

The following example illustrates the technique, consider two sentences to be aligned and to avoid language confusion let's assume symbolic sentences by denoting:

- Wsi: content words in the source sentence.
- NEsi: the Named Entity source words.
- Wti: the content words in the target sentence.
- NEti: the Named Entity target words.

The source and target sentences are represented as follows:

Source: *Ws1 Ws2 NEs1 NEs2 Ws3 Ws4 Ws5*
Target: *Wt1 Wt2 Wt3 NEt1 NEt2 NEt3 Wt4 NEt4*

After the first phase is applied, the remaining not aligned words might look like that:

Source: *NEs1 NEs2 Ws4 Ws5*
Target: *Wt3 NEt1 NEt2 NEt3 NEt4*

The example clarify that the elimination of some content words facilitates the task of NEs alignment since many of the words that might lead to confusion have been eliminated.

As shown in the above example, different mismatched identification of NEs could result from different identifiers. The "Multi-cost Named Entity Alignment by Content Words Elimination" technique helps in reducing alignment errors due to identification errors by reducing the candidate words for alignment and thus reducing the aligner confusion.

4.2 Phrase Based Named Entity Translation

For phrase-based NE translation, we used an approach similar to that presented by Tillman (Tillmann, 2003) for block generation with modifications suitable for NE phrase extraction. A block is defined to be any pair of source and target phrases. This approach starts from a word alignment generated by HMM Viterbi training (Vogel et. Al, 1996), which is done in both directions between source and target. The intersection of the two alignments is considered a high precision alignment and the union is considered a low precision alignment. The high precision alignments are used to generate high precision blocks which are further expanded using low precision alignments. The reader is referred to (Tillmann, 2003) for detailed description of the algorithm.

In our approach, for extracting NE blocks, we limited high precision alignments to NE phrases of the same NE types. In the expansion phase, the multi-cost function described earlier is used. Thus the blocks are expanded based on a cost depending on the type matching cost, the edit distance cost and the translation probability cost.

To explain this procedure, consider the following sentences pair:

وقد استدعى وزير الخارجية الياباني نوبوتاكا ماشيمورا السفير الصيني وانج يي "

"wqd AstdEY wzyr AlxArjyp AlyAbAny nwbwtAkA mA$ymwrA Alsfyr AlSyny wAnj yy"

"Japanese Foreign Minister Nobutaka Machimura has summoned the Chinese ambassador Wang Yee

The underlined words are the words that have been identified by the NE identifiers as person names. In the Arabic sentence, the identifier missed the second name of the first Named Entity (*mA$ymwrA)* and did not identify the word as person name by mistake. The high precision block generation technique will generate the following two blocks:

نوبوتاكا*(nwbwtAkA): Nobutaka*
وانج يي*: (wAnj yy) : Wang Yee*

The expansion technique will try to expand the blocks on all the four possible dimensions (right and left of the blocks in the target and source) of each block. The result of the expansion will be:

ماشيمورا نوبوتاكا *(nwbwtAkA mA$ymwrA) : Nobutaka Machimura*

Therefore, the multi-cost expansion technique enables expansions sensitive to the translation probability and the edit distance and providing a mechanism to overcome NE identifiers errors.

4.3 Named Entity Transliteration

NE transliteration is essential for translating Out Of Vocabulary (OOV) words that are not covered by the word or phrase based models. As mentioned earlier, phonetic and orthographic differences between Arabic and English make NE transliteration challenging.

We used a block based transliteration method, which transliterates sequence of letters from the source language to sequence of letters in the target language. These source and target sequences construct the blocks which enables the modeling of vowels insertion. For example, consider Arabic name "شكري $kry," which is transliterated as "Shoukry." The system tries to model bi-grams from the source language to n-grams in the target language as follows:

$k → shouk
kr→ kr
ry → ry

To obtain these block translation probabilities, we use the translation matrix, generated in section 4.1 from the word based alignment models. First, the translation matrix is filtered out to only preserve highly confident translations; translations with probabilities less than a certain threshold are filtered out. Secondly, the resulting high confident translations are further refined by calculating phonetic based edit distance between both romanized Arabic and English names. Name pairs with an edit distance greater than a predefined threshold are also filtered out. The remaining highly confident name pairs are used to train a letter to letter translation matrix using HMM Viterbi training (Vogel et al., 1996).

Each bi-gram of letters on the source side is aligned to an n-gram of letters sequence on the target side, such that vowels have very low cost to be aligned to NULL. The block probabilities are calculated and refined iteratively for each source and target sequences. Finally, for a source block *s* and a target block *t*, the probability of *s* being translated as *t* is the ratio of their co-occurrence and total source occurrence:

$$P(t \mid s) = N(t,s)/N(s).$$

The resulting block translation probabilities and the letter to letter translation probabilities are combined to construct a Weighted Finite State Transducer (WFST) for translating any source sequence to a target sequence.

Furthermore, the constructed translation WFST is composed with two language models (LM) transducers namely a letter trigram model and a word unigram model. The trigram letter based LM acts to provide high recall results while the word based unigram LM acts for providing high precisin results.

4.4 System Integration and Decoding

The three constructed models in the steps above, namely phrase-based NE translation, word-based translation, and transliteration, are used to generate hypotheses for each source NE phrase. We used a dynamic programming beam search decoder similar to the decoder described by Tillman (Tillmann, 2003).

We employed two language models that were built from NE phrases extracted from monolingual target data for each NE category under consideration. The first language model is a trigram language model on NE phrases. The second language model is a class based language model with a class for unknown NEs. Every NE that do exist in the monolingual data but out of the vocabulary of the phrase and word translation models are considered unknown. This helps in correctly scoring OOV hypothesis produced by the transliteration module.

5 Experimental Setup

We test our system for Arabic to English NE translation for three NE categories, namely names of persons, organizations, and locations. The system has been trained on a news domain parallel corpus containing 2.8 million Arabic words and 3.4 million words. Monolingual English data was annotated with NE types and the extracted named entities were used to train the various language models described earlier.

We manually constructed a test set as follows:

Category	No. of Phrases	No. of Words
Person names	803	1749
Organization names	312	867
Location names	345	614

The BLEU score (Papineni et al., 2002) with a single reference translation was deployed for evaluation. BLEU-3 which uses up to 3-grams is deployed since three words phrase is a reasonable length for various NE types. Table 1 reports the results for person names; the baseline system is a general-purpose machine translation system with relatively good Bleu score.

System	Bleu Score
Base line	0.2942
Word based only	0.3254
Word + Phrase	0.4620
Word + Phrase + Transliteration	0.5432

Table 1: Person Names Results

Table 2 reports the bleu score for Location category with the same three systems presented before with persons:

System	Bleu Score
Base line	0.2445
Word based only	0.3426
Word + Phrase	0.4721
Word + Phrase + Transliteration	0.4983

Table 2: Locations Names Results

Table 3 reports the bleu score for Organization category with the same three systems presented before:

System	Bleu Score
Base line	0.2235
Word based only	0.2541
Word + Phrase	0.3789
Word + Phrase + Transliteration	0.3876

Table 3: Organizations Names Results

Figure 1, illustrates various BLEU scores for various categories. The results indicate that phrase based translation provided enhancement for all NE types, while transliteration proved more effective for person names.

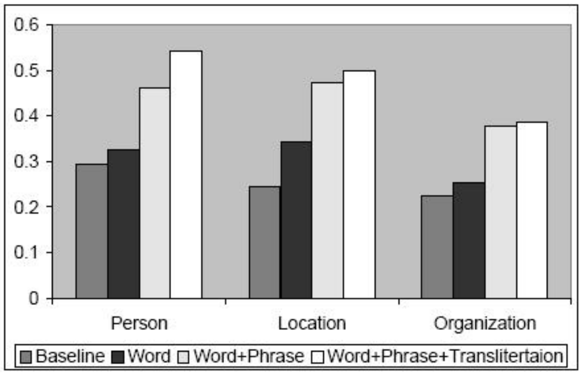

Figure 1: Various BLEU scores for various categories

It is also worth mentioning that evaluating the system using a single reference has limitations; many good translations are considered wrong because they do not exist in the single reference.

6 Conclusion and Future Work

We have presented an integrated system that can handle various NE categories and requires the regular parallel and monolingual corpora which are typically used in the training of any statistical machine translation system along with NEs identifier. The proposed approach does not require any costly special resources, lexicons or any type of annotated data.

The system is composed of multiple translation modules that give flexibility for different named entities type’s translation requirements. This gives a great flexibility that enables the system to handle NEs of any type.

We will evaluate the effect of the system on CLIR and MT tasks. We will also try to investigate new approaches for deploying NE translation in general phrase based MT system.

Acknowledgment

We would like to thank Salim Roukos and Kishore Papineni for several invaluable suggestions and guidance. We would like also to thank Christoph Tillmann for help with various components.
We would like also to thank Kareem Darwish for his invaluable help in editing this paper.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Machine Transliteration of Names in Arabic Text. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263–311.

Radu Florian, Hany Hassan, Abraham Ittycheriah, H. Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, Salim Roukos: A Statistical Model for Multilingual Entity Detection and Tracking. HLT-NAACL 2004: 1-8

Fei Huang, Stephan Vogel and Alex Waibel, Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization, in the Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual and Mixed-language Named Entity Recognition, July, 2003

Leah Larkey, Nasreen AbdulJaleel, and Margaret Connell, What's in a Name? Proper Names in Arabic Cross-Language Information Retrieval. CIIR Technical Report, IR-278,2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of machine translation. In Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02), pages 311–318, Philadelphia, PA, July.

Bonnie G. Stalls and Kevin Knight.. Translating Names and Technical Terms in Arabic Text. In Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages. 1998.

Christoph Tillmann,. A Projection Extension Algorithm for Statistical Machine Translation. In Proc of Empirical Methods in Natural Language Processing, 2003

Stefan Vogel, Hermann Ney, and Christoph Tillmann.. HMM Based Word Alignment in Statistical Machine Translation. In Proc. of the 16th Int. Conf. on Computational Linguistics (COLING), 1996

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