King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Transliteration normalization for Information Extraction and Machine Translation

Yuval Marton, Imed Zitouni *

Microsoft, Bellevue, WA, United States

KEYWORDS

Arabic;
Named Entity Recognition;
Transliteration;
Name normalization;
Information Extraction;
Machine Translation

Abstract Foreign name transliterations typically include multiple spelling variants. These variants cause data sparseness and inconsistency problems, increase the Out-of-Vocabulary (OOV) rate, and present challenges for Machine Translation, Information Extraction and other natural language processing (NLP) tasks. This work aims to identify and cluster name spelling variants using a Statistical Machine Translation method: word alignment. The variants are identified by being aligned to the same "pivot" name in another language (the source-language in Machine Translation settings). Based on word-to-word translation and transliteration probabilities, as well as the string edit distance metric, names with similar spellings in the target language are clustered and then normalized to a canonical form. With this approach, tens of thousands of high-precision name transliteration spelling variants are extracted from sentence-aligned bilingual corpora in Arabic and English (in both languages). When these normalized name spelling variants are applied to Information Extraction tasks, improvements over strong baseline systems are observed. When applied to Machine Translation tasks, a large improvement potential is shown.

1. Introduction

Foreign names typically have multiple spelling variants after translation or transliteration (where translation aims to preserve meaning, while transliteration aims to preserve sound, given differences in the languages' sounds and writing systems). These spelling variants present challenges for many natural language processing (NLP) tasks, as they increase both the vocabulary size and Out-of-Vocabulary (OOV) rate,[1] exacerbate the data sparseness problem, and may introduce inconsistencies (in spelling or in reference as multiple entities). When different spelling variants are generated for the same name in one document, it reduces the named entity resolution scores and the readability of Machine Translation output. This paper addresses this problem by replacing each spelling variant with a corresponding canonical form. Such text normalization could potentially benefit many NLP tasks, including information retrieval, Information Extraction, question answering, speech recognition, and Machine Translation.

* Corresponding author.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

[1] OOV rate: how often the model processes an input term that it has not been trained on. Typically, models perform poorly on OOV terms, whether they be Machine Translation, parsing, Mention Detection or other NLP models.

http://dx.doi.org/10.1016/j.jksuci.2014.06.011
1319-1578 © 2014 Production and hosting by Elsevier B.V. on behalf of King Saud University.

Name spelling variants have been studied mostly in Information Retrieval (IR) research, especially in query expansion and cross-lingual IR. Bhagat and Hovy (2007) proposed two approaches for (primarily English) spelling variant generation, based on letters-to-phonemes mapping and the SoundEx algorithm (Knuth, 1973). Raghavan and Allan (2005) proposed several techniques to group names in Automatic Speech Recognition (ASR) output and evaluated their effectiveness in spoken document retrieval (SDR). Both approaches use a named entity extraction system to automatically identify names. For multi-lingual name spelling variants, Linden (2006) proposed using a general edit distance metric with a weighted FST to find technical term translations (which were referred to as "cross-lingual spelling variants"). These variants are typically translated words with similar stems in another language. Toivonen and colleagues (2005) proposed a two-step fuzzy translation technique to solve similar problems. Al-Onaizan and Knight (2002), Huang et al. (2003), and Ji and Grishman (2007) investigated the general name entity translation problem, especially within the context of Machine Translation.

All of these approaches rely on name taggers and other classifiers to directly identify the variants. This work, however, aims to identify name spelling variants using *crosslingual* information, with application to Arabic and English. Instead of using a named entity tagger to directly identify names and their spelling variants, we link spelling variants with a name in another language via a method that is widely used in Statistical Machine Translation: word alignment. From sentence-aligned bilingual corpora, we collect word co-occurrence statistics and calculate word translation probabilities (including transliterated words).[2] For each source-side word, we group its target-side aligned counterparts into clusters according to target-side string edit distances. Then, we calculate the transliteration cost between the source word and each target-side cluster (see Section 3). Word pairs with small transliteration costs are considered name variants. We then normalize all names in each cluster to the most frequent form.

Note that spelling variation does not necessarily stem from transliteration or translation, e.g.,

- Cindy and Cyndi
- Kacey and KC (read-aloud initials)
- Cl8n and Clayton (informal communication writing style)
- Dialectal differences (e.g., الجزيرة vs. الڤزيرة)

However, these other cases most likely should not be clustered and normalized (except, perhaps, the informal writing style), as they are likely to refer to different people/entities. These cases are outside the scope of this work.

We applied our approach to extract name transliteration spelling variants from bilingual Arabic–English corpora. We obtained tens of thousands of high-precision name translation pairs. We further applied these spelling variants to Machine Translation (MT) and Information Extraction (IE) tasks, and observed a statistically significant improvement over a strong baseline on the IE task, and a close to "oracle" improvement on a small test set on the MT task.

After an Arabic-focused survey of related work (Section 2), we describe our model setting in both Information Retrieval and Statistical Machine Translation (Section 3). We then detail our past and new experiments (Section 4). We follow up with an analysis of the results (Section 4) and conclude with possible future work (Section 5).

2. Related work

In addition to the work we mentioned earlier, there has been much related work in both IE and MT. We focus here on Arabic (or Arabic and English) related work.

The idea of using cross-language propagation to boost performance has been applied by several researchers. For example, Tackstromand et al. (2012) show how the use of cross-lingual word clusters for the transfer of linguistic structure improves system performance. Other research studies (such as Goldsmith, 2001; McCallum and Nigram, 1998; Yarowsky, 1995) report the use of cross-language propagation to boost the performance of different systems, namely, morphological segmentation, text categorization and word segmentation, respectively. These approaches are based on monolingual data. Rogati et al. (2003) use a Statistical Machine Translation (SMT) system to build an Arabic stemmer. The obtained stemmer has a performance of 87.5%. Ide et al. (2002) use the aligned versions of George Orwell's "Nineteen Eighty-Four" in seven languages to determine sense distinctions that can be used in the Word Sense Disambiguation (WSD) task. They report that the automatically obtained tags are at least as reliable as the tags created by human annotators. Zitouni et al. (2005) attempt to enhance a Mention Detection model of a foreign language by using an English Mention Detection system. They used an SMT system to (i) translate the text into English, (ii) run the English model on the translated text, and (iii) propagate the outcome to the original text. Das and Petrov (2011) try a similar approach but apply it to POS taggers. Both approaches require an SMT system.

The detection (or generation) of named entity variants has also been explored and evaluated in SMT, often as a subset of a paraphrase generation task. In this case, variants (paraphrases) are used to augment translation tables that are missing the variants, unlike our work, which uses them for the normalization of existing terms. Hereafter, we call the term to be paraphrased the *anchor*.

Callison-Burch et al. (2006) proposed a general paraphrasing method by "pivoting" through additional languages in SMT tables and back to the original language. The method is as follows: for each anchor, find its translation(s) in the table and "pivot" through each translation term back to the original (the anchor's) language, i.e., translate back. The back-translations are often good paraphrases and potentially good name variants. Our work uses similar pivoting but then further clusters terms by edit distance and transliteration cost. Interestingly, Callison-Burch et al. (2006) excluded named entities from their experiments, presumably due to noisier results in this particular subset problem. Callison-Burch (2008) improved this method with syntactic constraints. Many publications used or extended the pivoting method, some of which we list below. While the

[2] Throughout this article, we sometimes use the term 'translation' loosely, encompassing both translation and transliteration, as there is no explicit representational difference between the two in Statistical Machine Translation phrase tables.

above-mentioned pivot approach paraphrased the source language in translation tasks, other variants used the pivot method to paraphrase the target language, thus creating additional (pseudo) reference translations to improve parameter estimation and increase translation coverage (Madnani et al., 2007, 2008; Madnani and Dorr, 2013).

Another paraphrasing method is *distributional paraphrasing*. This method relies on the Distributional Hypothesis (Harris, 1954; Firth, 1957), which assumes that words similar in meaning keep similar company. Distributional paraphrasing has been applied in the context of IR (Pasca and Dienes, 2005; Bhagat and Ravichandran, 2008) and SMT (Marton et al., 2009; Marton, 2010, 2013). The latter line of work (Marton et al., 2009; Marton, 2010; 2013) introduced a variant that can paraphrase anchors of arbitrary length on-the-fly (without having to pre-compute an entire collocation matrix). It included the paraphrasing of named entities, pointing out weaknesses of both pivot and distributional methods in paraphrasing named entities. A hybrid pivot-distributional approach generates paraphrases bilingually (using the pivot method) and then re-ranks the paraphrase candidates using distributional similarity (Chan et al., 2011).

A third approach relies on crowd-sourcing. Denkowski et al. (2010) used the crowd-sourcing platform Amazon Mechanical Turk to generate paraphrases for an English–Arabic translation task. Spelling variants were handled as a subset task of translation from Arabic dialects to English (Zbib et al., 2012), also through crowd-sourcing.

Li et al. (2013) proposed a ‘‘Name-aware Machine Translation’’ approach that jointly annotates parallel corpora and extracts name-aware translation rules during decoding. They also proposed a weighted automated MT quality scoring, giving more weight to content words and words of importance (potentially names).

Abdel Fattah and Ren (2008) present several approaches for the extraction of transliteration proper-noun pairs from Arabic–English parallel corpora based on different similarity measures between the English and Romanized Arabic proper nouns under consideration. The strength of the proposed techniques is their effectiveness on low-frequency proper noun pairs. The reader can also refer to the book by Izwaini (2011) for additional approaches to transliterations.

As mentioned above, none of these approaches cluster name variants as we do or use them for name normalization. The work presented here shares many similarities with that of Huang et al. (2008); in this work, we too view the problem from a paraphrasing angle.

3. Detection and clustering of transliterated name entity variants

Our approach requires parallel data, where each word in the source-side language (e.g., Arabic) is aligned to one or many words in the target-side language (e.g., English) with a translation probability (including transliteration cases). The translation probability can be obtained through a translation model similar to the one described in Vogel et al., 1996; Ge, 2004.

An example of entry is shown in Table 1 for the Arabic word ‘‘الخروب |Alxrwb’’, where we show several possible translation candidates (including both transliteration variants and English words), all of which are actual entries taken from our model. Because the lexical translation probabilities are distributed among these variants, none of them has the highest probability compared to all translation candidates. As a result, the incorrect translation, *iqlim*, is assigned the highest probability, and hence, it is often selected in Machine Translation output. To fix this problem, we propose identifying and grouping the target spelling variants from among all translation candidates, converting them into a canonical form and merging their translation scores.

For each source word or phrase in the source language that has alignment, we cluster its target translations/transliterations based on string edit distances using the group average agglomerative clustering algorithm (Manning and Schütze, 1999). Initially, each target word or phrase is in a single cluster. We calculate the average editing distance between any two clusters and merge them if the distance is smaller than a certain threshold. This process repeats until the minimum distance between any two clusters is above a certain threshold. In the above example, *alkharrub, al-kharub, al-khurub* and *al-kharroub* are grouped into a single cluster, and each of the other words remains in its own single word cluster.

Note that the source word may not be a name, while its translations may still have similar spellings. An example is the Arabic word أعلم (AElm), which is aligned to the English words *brief, briefing, briefed* and *briefings*. To detect whether a source word is a name, we calculate the *transliteration cost* between the source word and its target translation cluster, which is defined as the average phonetic distance between the source word and each target word in the cluster. Source words whose transliteration cost is lower than an empirically selected threshold are considered as names, based on the assumption that names are typically transliterated, and transliteration aims to preserve sound by definition. Conversely, source words with high transliteration cost are considered non-names, assuming only names would be transliterated, while other words would be translated (i.e., preserving meaning but not sound). As non-names, they are therefore not used for linking spelling variants in the target language.

The phonetic distance between a source word and a target word is calculated based on the character transliteration model, which can be trained from bilingual name transliteration pairs. We segment the source and target names into characters and then run HMM alignment on the source and target character pairs (Vogel et al., 1996; Ge, 2004). After the training, character transliteration probabilities can be estimated from the relevant frequencies of character alignments.

Suppose the source word f contains m characters, $f_1, f_2, \ldots, f_m$, and the target word e contains n characters, $e_1, e_2, \ldots, e_n$. For $j = 1, 2, \ldots, n$, letter e_j is aligned to character f_{aj} according to the HMM aligner. Under the (naïve) assumption that character alignments are independent, the word transliteration probability is calculated as

Table 1 English translations of a Romanized Arabic name *Alxrwb* with translation probabilities.

الخروب |Alxrwb							
Iqlim [0.22]	al-kharrub [0.16]	al-kharub [0.11]	Overflew [0.09]	junbulat [0.05]	al-khurub [0.05]	Hours [0.04]	al-kharroub [0.03]

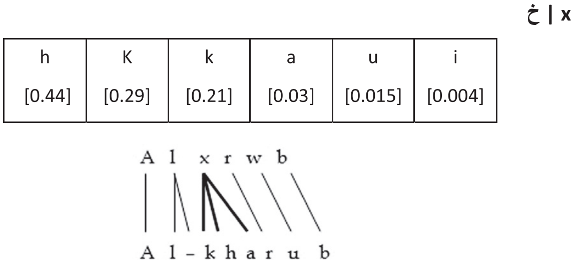

Figure 1 Example of the learned A–E character transliteration model with probabilities and its application in the alignment between a Romanized Arabic name and an English translation.

$$P(e|f) = \prod_{j=1}^{n} p(e_j|f_{a_j}) \tag{2.1}$$

where $p(e_j|f_{aj})$ is the character transliteration probability. Note that in the above configuration, one target character can be aligned to only one source character, and one source character can be aligned to multiple target characters. Note that the character transliteration probability is a phonetic similarity measure, i.e., its inverse (or log inverse) is the phonetic distance measure.

An example of the trained Arabic-to-English character transliteration model is shown in Fig. 1. The Arabic character خ (*x*) is aligned with high probabilities to English letters with the most similar pronunciation (*kh*). Because the written form of Arabic words often omits vowel markers, English vowels are also aligned to Arabic characters (e.g., *x–kha*). Given this model, the characters within a Romanized Arabic name and its English transliteration are aligned as shown in Fig. 1. Note that this model does not commit to any specific linguistic theory or representation, nor does it claim to be 100% correct; it merely represents the statistically learned character alignments. The Arabic alphabet includes 28 characters (and gemination and vowel diacritics), and the English alphabet includes 52 letters (26 lowercase letters and 26 uppercase letters).

4. Model setting

4.1. Application to Machine Translation

One important application of detecting and normalizing name translation spelling variants is to boost the performance of a Machine Translation system. Given the name spelling variants, we updated both the translation model (Section 4.1.1) and the language model (Section 4.1.2), transferring the variants' probabilities to the canonical form. In addition to improving spelling consistency, it also often helps the correct translation (regardless of spelling variation) win over other translation candidates.

The SMT decoder used for our baseline is a phrase-based decoder similar to the one in Al-Onaizan and Papineni, 2006. Given a source sentence, the decoder tries to find the translation hypothesis with minimum translation cost, which is defined as the log-linear combination of different feature functions, such as the translation model cost, language model cost, distortion cost and sentence length cost. The translation cost includes word (lexical) translation probabilities and phrase translation probabilities.

4.1.1. Updating the translation model

Updating lexical translation probabilities: Given target-side name spelling variants $\{t_1, t_2, \ldots, t_m\}$ for a source-side name s, assume without loss of generality that $t_1, t_2, \ldots, t_m$ are sorted by their lexical translation probabilities, $p\ (t_1|s) \geqslant p\ (t_2|s) \geqslant \ldots \geqslant p\ (t_m|s)$.

We select t_1 as the canonical spelling, and add to its probability score all other spelling variants' translation probabilities:

$$p(t_1|S) \leftarrow \sum_{j=1}^{m} p(t_j|S).$$

These other (non-canonical) spelling variants are then assigned zero probability. Table 2 shows the updated word translation probabilities for “الخروب |Alxrwb”. Compared with Fig. 1, the translation probabilities from several spelling variants are aggregated and merged into the canonical form, *al-kharrub*, which now has the highest probability in the new model.

Updating phrase translation probabilities: The phrase translation table includes source phrases, their target phrase translations and the frequencies of the bilingual phrase pair alignment. The phrase translation probabilities are calculated based on their alignment frequencies, which are collected from word-aligned parallel data. To update the phrase translation table, for each phrase pair, including a name in the source phrase and its spelling variant in the target phrase, we replace the target-side name with its canonical spelling. After the mapping, two target phrases, differing only in spelling variants of names, may end up identical after normalization to the canonical form, and their alignment frequencies would be added together. Phrase translation probabilities are then re-estimated with the updated alignment frequencies. The effect would be similar to what is illustrated in Table 2 (except that each translation unit may be longer than a single token).

4.1.2. Updating the language model

Language modeling is one of the key components of an SMT decoder in delivering a well-formed output. Because the updated translation model can produce only the canonical form of a group of spelling variants, the language model should be updated so that all m-grams ($1 \leqslant m \leqslant N$) containing spelling variants of each other are normalized (and their counts added), resulting in the canonical form of the m-gram.

Table 2 English translations of an Arabic name الخروب|*Alxrwb* with the updated word translation model scores in the second line. “Winner” translations are in bold; non-canonical spelling variants are in *italics*.

الخروب|Alxrwb								
Orig.:	**Iqlim [0.22]**	al-kharrub [0.16]	al-kharub [0.11]	overflew [0.09]	junbulat [0.05]	al-khurub [0.05]	Hours [0.04]	al-kharroub [0.03]
Updated:	Iqlim [0.22]	**al-kharrub [0.35]**	*al-kharub [0.0]*	overflew [0.09]	junbulat [0.05]	*al-khurub [0.0]*	Hours [0.04]	*al-kharroub [0.0]*

Please cite this article in press as: Marton, Y., Zitouni, I. Transliteration normalization for Information Extraction and Machine Translation. Journal of King Saud University – Computer and Information Sciences (2014), http://dx.doi.org/10.1016/j.jksuci.2014.06.011

Two m-grams are considered spelling variants of each other if they contain words t_1^i t_2^i ($t_1^i \neq t_2^i$) at the same position i in the m-gram and if t_1^i and t_2^i belong to the same spelling variant group, as defined in Section 3.

An easy way to achieve this update is to replace every spelling variant in the original language model training data with its corresponding canonical form, and then rebuild the language model. However, because we do not want to replace words that are not names, we need to have a mechanism for detecting names. For simplicity, in our experiments (with English language models), we assumed a word is a name if it is capitalized, and we replaced spelling variants with their canonical forms only for words that start with a capital letter. Experimental results are reported in Section 5.

4.2. Application to Information Extraction

Information Extraction is a crucial step toward understanding a text, as it identifies the important conceptual objects in a discourse. We address here one important and basic task of Information Extraction: *mention detection*.[3] We call instances of textual references to objects *mentions*, which can be named (e.g., *John Smith*), nominal (*the president*) or pronominal (e.g., *he*, *she*). For instance, in the sentence

Queen Rania Al Abdullah said she has no comments.

there are three mentions: *Queen*, *Rania Al Abdullah* and *she*, all of which refer to the same entity. Similar to many classical NLP tasks, we formulate the mention detection problem as a classification problem by assigning a label to each token in the text, indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. Good performance in many natural language processing tasks has been shown to depend heavily on integrating many sources of information (Florian et al., 2004). We select an exponential classifier, the Maximum Entropy (MaxEnt henceforth) classifier, which can integrate arbitrary types of information and make a classification decision by aggregating all information available for a given classification (Berger et al., 1996). In this paper, the MaxEnt model is trained using the *sequential conditional generalized iterative scaling* (SCGIS) technique (Goodman, 2002), and it uses a *Gaussian prior* for regularization (Chen and Rosenfeld, 2000).

In ACE, there are seven possible mention types: person, organization, location, facility, geopolitical entity (GPE), weapon, and vehicle. Experiments are run on Arabic and English. Our baseline system achieved very competitive result among systems participating in the ACE 2007 evaluation. It uses a large range of features, including lexical and syntactic features, and the output of other Information Extraction models. These features were described in Zitouni and Florian (2008) and Florian et al. (2004) and are not discussed here.

4.2.1. Canonical form of name-spelling-variants as a feature

We focus here on examining the effectiveness of name spelling variants in improving mention detection systems. To do so, we created a new feature where for each input token x, we fire its canonical form i (class label) $C(x_i)$, which is representative of name spelling variants of x_i. This name spelling variant feature is also used in *conjunction* with the lexical features (e.g., words and morphemes in a 3-word window, prefixes and suffixes of length up to 4, stems in a 4-word window for Arabic) and syntactic features (e.g., POS tags, text chunks).

4.2.2. Propagation of mentions as a feature

Another approach in Information Extraction is to study the effectiveness of mention-detection and name-entity-recognition by propagating it from English to Arabic. Our goal is to not be limited to name-spelling-variants but to use all mentions in a resource-rich-language, English in our case, in order to improve Arabic Information Extraction.

Our approach requires word alignment and a mention detection system trained on English. The first step consists of running the mention-detection system on English training data, resulting in a tagged text. We then group mentions by class in different dictionaries. During the decoding of Arabic text, when we encounter a token or a sequence of tokens that is an entry in a dictionary, we fire its corresponding class; the feature is fired only when we find an exact match between sequences of tokens (including single tokens) in the text and in the dictionary.

5. Experiments

5.1. Name spelling variant precision

We extracted Arabic-to-English and English-to-Arabic name translation variants (including transliterations) from sentence-aligned parallel corpora released by LDC.[4] The Arabic–English parallel corpora include 5.6 M sentence pairs, 845 K unique Arabic words and 403 K unique English words. We trained a word translation model by running HMM alignment on the parallel data, grouping target translation with similar spellings and computing the average transliteration cost between the Arabic word and each English word in the translation clusters according to Formula (2.1). We sorted the name translation groups according to their transliteration costs, and selected 300 samples at different ranking positions for evaluation (20 samples at each ranking position). The quality of the name translation variants was judged as follows: for each candidate name translation group $\{t_1, t_2, \ldots, t_m|s\}$, if the source word s is a name and all of the target spelling variants are correct transliterations, it receives a credit of 1. If s is not a name, the credit is 0. If s is a name but only part of the target spelling variants are correct, it receives partial credit n/m, where n is the number of correct target translations. We evaluated only the precision of the extracted spelling variants,[5] as judged by proficient Arabic speakers. As seen in Fig. 2, the precision of the top 22 K Arabic–English name translations was 96.9%. Among them 98.5% of the automatically aligned Arabic words were names. The precision decreases as more non-name Arabic words are included. On average, each Arabic name has 2.47 English spelling variants, although there are some names with more than 10 spelling variants.

[3] We adopt here the ACE (NIST, 2007) nomenclature.

[4] Linguistic Data Consortium at https://www.ldc.upenn.edu.

[5] Evaluating recall requires one to manually look through the space of all possible transliterations (hundreds of thousands of entries), which is impractical.

Please cite this article in press as: Marton, Y., Zitouni, I. Transliteration normalization for Information Extraction and Machine Translation. Journal of King Saud University – Computer and Information Sciences (2014), http://dx.doi.org/10.1016/j.jksuci.2014.06.011

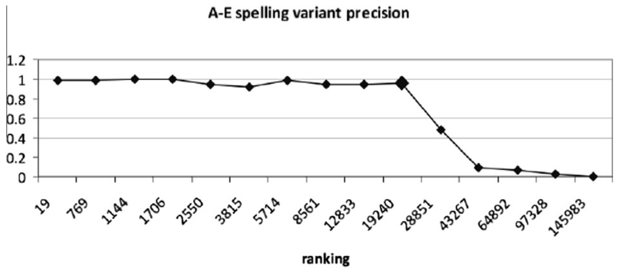

Figure 2 Arabic–English (A–E) name spelling variant precision curve. (Precision of evaluation sample at different ranking positions. The larger square indicates the cutoff point).

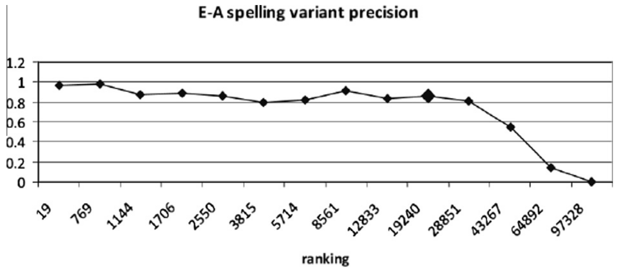

Figure 3 English–Arabic (E–A) name spelling variant precision curve.

Switching the source and target languages, we obtained English–Arabic name spelling variants, i.e., one English name with multiple Arabic spellings. As seen in Fig. 3, the top 20 K English–Arabic name pairs are obtained with a precision above 87.9%, and each English name has 3.3 Arabic spellings on average. Table 3 shows some Arabic–English and English–Arabic name spelling variants, where Arabic words are represented in their Romanized form.

5.2. Machine Translation

Table 4 English translation output with the baseline SMT system and the system with updated models.

Source	Alm&tmr AlAwl lAqlym *Alxrwb* AlErby AlmqAwm
Reference	the first conference of the Arab resistance in Iqlim *Kharoub*
Baseline	the first conference of the Arab *regional* resistance
Updated model	first conference of the *Al-Kharrub* the Arab resistance

We applied the Arabic-to-English name spelling variants to a Machine Translation task. Our baseline system was trained with 5.6 M Arabic–English sentence pairs, using the same training data used to extract the Arabic-to-English spelling variants. The language model was a five-gram model with modified Kneser–Ney smoothing, trained on approximately 3.5 billion words. After pruning (using count cutoffs), the model contained a total of 935 million *N*-grams. We then updated the translation models and the language model with the name spelling canonical forms and updated their probabilities and scores accordingly.

Table 4 shows a Romanized Arabic sentence, the translation output from the baseline system and the output from the updated models. In the baseline system output, the Arabic name “Alxrwb” was incorrectly translated into “regional”. This error was fixed in the updated model, where both the translation and language models assign higher probabilities to the correct translation “al-kharroub” after spelling variant normalization.

We evaluated the updated SMT models on a test set including 70 documents: 42 newswire documents and 28 weblog and newsgroup documents. There are 669 sentences with 16.3 K Arabic words in the test data. The results were evaluated against one reference human translation using BLEU (Papineni et al., 2001) and TER (Snover et al., 2006) scores. The results using the baseline decoder and the updated models are shown in Table 5. Applying the updated language model (ULM) and the translation model (UTM) led to a small reduction in TER. Additionally applying similar name spelling normalization to the reference translation resulted in a BLEU score increase of 0.1 points and a TER reduction of nearly 0.3 points. We discuss potential reasons for the lack of significant gains in Section 5.

5.3. Information Extraction

Similarly to classical NLP tasks, such as text chunking (Ramshaw and Marcus, 1994) and named entity recognition (Tjong Kim Sang, 2002), we formulate mention detection as a sequence classification problem by assigning a label to each token in the text, indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. For Arabic, blank-delimited words are composed of zero or more prefixes, followed by a stem and zero or more suffixes.

Table 3 Arabic-to-English and English-to-Arabic name spelling variant examples. *Italic words* represent different persons with similarly spelled names.

Lang. pair	Source name	Target spelling variants
Arabic to English	Alxmyny	khomeini al-khomeini al-khomeni khomeni khomeyni *khamenei khameneh'i*
	krwby	karroubi karrubi krobi karubi karoubi kroubi
	gbryAl	gabriel gabrielle gabrial ghobrial ghybrial
English to Arabic	cirebon	syrybwn syrbwn syrbn kyrybwn bsyrybwn bsyrwbwn
	mbinda	mbyndA mbndA mbydA AmbyndA AmbAndA mbynydA
	nguyen	njwyn ngwyn ngwyyn ngyyn Angwyn nygwyyn nygwyn wnjwyn njwyyn nyjyn bnjwyn wngyyn ngwyAn njyn nykwyn

Please cite this article in press as: Marton, Y., Zitouni, I. Transliteration normalization for Information Extraction and Machine Translation. Journal of King Saud University – Computer and Information Sciences (2014), http://dx.doi.org/10.1016/j.jksuci.2014.06.011

Table 5 MT scores with updated TM and LM.

	BLEU r1n4	TER
Baseline	27.1	51.7
Baseline + ULM + UTM	27.2	51.5
Ref. Normalization	27.2	51.4

Table 6 Performance of English and Arabic mention detection systems with and without the use of name spelling variants (NSV) and gazetteers (Gaz). Performance is presented in terms of Precision (P), Recall (R) and F-measure (F).

	Baseline			Baseline + NSV			Baseline + Gaz		
	P	R	F	P	R	F	P	R	F
English	84.4	80.6	82.4	84.6	80.9	82.7	–	–	–
Arabic	81.8	71.7	76.4	82.6	73.9	78.0	82.3	72.8	77.26

Each prefix, stem or suffix is a token, and any contiguous sequence of tokens can represent a mention. We decided to "condition" the output of the system on the segmented data: the text is segmented first into tokens, and classification is then performed on tokens. The segmentation model is similar to that presented by Lee et al. (2003) and obtains an accuracy of 98%.

The classification is performed with a statistical approach built around the maximum entropy (MaxEnt) principle (Berger et al., 1996), which has the advantage of combining arbitrary types of information in making a classification decision. The mention detection system tags each token x_i in a sentence $x_1, \ldots, x_n$ with a label y_i as follows:

- if it is not part of any entity, $y_i = \mathrm{O}$ (O for "outside any mentions")
- if it is part of an entity, it is composed of a subtag specifying whether it starts a mention (*B*-) or is inside a mention (*I*-), and another sub-type corresponding to the mention type (e.g., *B-PERSON*).

Mention detection system experiments are conducted on the ACE 2007 data sets in Arabic and English (NIST, 2007). There are seven possible types: person, organization, location, facility, geopolitical entity (GPE), weapon, and vehicle. For English mention detection systems, use the word forms as the tokens for classification.

Because the evaluation test set is not publicly available, we split the publicly available training corpus into an 85%/15% data split. To facilitate future comparisons with the work presented here, and to simulate a realistic scenario, the splits were created based on article dates: the test data are selected as the latest 15% of the data in chronological order. In this way, the documents in the training and test data sets do not overlap in time, and the content of the test data is more recent than the training data. For English, we use 499 documents for training and 100 documents for testing, while for Arabic, we use 323 documents for training and 56 documents for testing. English and Arabic mention detection systems use a large range of features, including lexical (e.g., words and morphs in a 3-word window, prefixes and suffixes up to a length of 4, stems in a 4-word window for Arabic). These features were described in Zitouni and Florian (2008) and Florian et al. (2004) in further detail.

Our goal here is to investigate the effectiveness of name spelling variant information as well as mention-propagation in improving the mention detection system's performance.

The results in Table 6 show that the use of name spelling variants (**NSVs**) indeed improves the mention detection system's performance. A sizeable improvement is obtained in recall – which is to be expected, given the method. Improvement is obtained also in precision, leading to systems with better performance in F-measure (82.4 vs. 82.7 for English and 76.4 vs. 77.26 for Arabic). In the case of Arabic, the use of mention-propagation to build gazetteer (**Gaz**) features further improves the performance. This was not the case for English, where the use of information propagation from Arabic did not change the performance. This is explained by the fact that English uses a richer set of resources and uses more training data than Arabic. Hence, it was not possible to benefit from gazetteers extracted from Arabic. On the other hand, the Arabic system uses only lexical features, and hence, it was possible to benefit from information propagation from the English language with a richer resource. When the Arabic system used a richer set of information, including syntactic information (POS tags, text chunks) and the output of other Information Extraction models, the baseline performance increased to 81.6F (84.3 Precision and 79.0 Recall). In this case, the use of name spelling variants led to a 0.1 improvement in Precision and 0.1 in Recall. To measure whether the improvement in performance of a particular approach over another is statistically significant, we use the stratified bootstrap re-sampling significance test (Noreen, 1989). This approach is used in the named entity recognition shared task of CoNLL-2002.[6] Based on this stratified bootstrap re-sampling approach, the small improvement in performance was shown to be statistically significant.

However, the small improvement obtained for Arabic is not statistically significant based on the approach described earlier. One hypothesis for this result is that Arabic name spelling variants are not rich enough and that a better tuning of the alignment score is required to improve precision.

6. Discussion

According to our error-analysis, the significant number of Arabic names observed in the parallel corpus, where many of them do not appear in the training corpus, has significantly helped the MT and IE models capture this new information and/or correct the type assigned. Some of the relevant examples in our data are the following: (i) the facility mention (mbnY blfwr – Belvoir Building), (ii) the GPE name (kAbwl – Kabul), and (iii) the person mention (AlbEvyyn – the Baathists). These mentions were only tagged correctly when we used our approach. In other words, the error-analysis clearly shows that one possible way to obtain further improvement is to increase the parallel data, and hence to increase the number of matches between (1) names that are wrongly tagged and (2) names in the target language in the parallel corpus. The second parameter can be indirectly increased by increasing the size of the parallel data. However, obtaining 10 or 20 times

[6] http://www.cnts.ua.ac.be/conll2002/ner/.

Please cite this article in press as: Marton, Y., Zitouni, I. Transliteration normalization for Information Extraction and Machine Translation. Journal of King Saud University – Computer and Information Sciences (2014), http://dx.doi.org/10.1016/j.jksuci.2014.06.011

more parallel data that is hand-aligned is expensive and requires several months of human/hours work. For this reason, one may opt to use an unsupervised approach by selecting a parallel corpus that is automatically aligned.

The results obtained by all of these experiments help answer an important question: when trying to improve mention detection systems in a resource-poor language, should we invest in building resources or should we use propagation from a resource-rich language to (at least) bootstrap the process? The cost-effective answer seems to be the latter. Having said that, we also note that the improvement in performance shows diminishing returns with increasing resource availability. While the evidence here is not definitive, this trend is expected.

Although the significance of correct name translation cannot be fully represented by BLEU and TER scores,[7] we still want to understand why the improvement in translation quality was so small. After further error analysis, we found that in our test-set, approximately 2.5% of the Arabic words are names with English spelling variants. Among them, 73% name spelling errors can be corrected with the translation spelling variants obtained in Section 5.1. However, because the SMT system was trained on the same bilingual data from which the name spelling variants were extracted, some of these Arabic names were already correctly translated in the baseline system. Thus, the room for improvement in this setting was small.

We followed this with an oracle experiment, manually correcting the name translation errors in the first 10 documents (89 sentences with 2545 words). With only six name translation errors corrected, this reduced the TER from 48.83 to 48.65. We take this result as supporting our assumption that our approach has the potential for a much larger impact.

7. Conclusion and future work

We presented an approach to detect name variants, with a focus on transliteration variants. Our approach uses a pivot translation: a name in another (source) language, which is aligned to all of the (target language) name variants. These variants are then clustered by edit distance and transliteration score. Finally, each cluster is normalized to the most frequent form. We applied our approach to Information Extraction and Machine Translation tasks in Arabic and English. We observed significant gains over a strong baseline in Information Extraction. We also saw a potential for substantial gains in Machine Translation.

In the future, we intend to extend this work to use semi-supervised and unsupervised approaches that can make use of cross-language information propagation to bootstrap the performance of both Information Extraction and Machine Translation. We also intend to expand the applicability of our approach to non-translated settings, in which it will be necessary to distinguish between similarly spelled names denoting separate entities (despite a short edit distance or small transliteration cost) and those denoting spelling variants of the same entity.

We believe it is important for the research community to continue to invest in building better resources in non-English "source" languages such as Arabic, as it seems to be the most promising approach. It is also our belief that using a cross-lingual propagation approach can help bootstrap the process.

[7] These scores treat information-bearing words, such as names, the same as any other tokens, such as punctuations.

References

Abdel Fattah, M., Ren, F., 2008. English–Arabic proper-noun transliteration-pairs creation. J. Am. Soc. Inf. Sci. Technol. 59 (10), 1675–1687. http://dx.doi.org/10.1002/asi.20877.

Al-Onaizan, Y., Papineni, K., 2006. Distortion models for statistical machine translation. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia.

Al-Onaizan, Y., Knight, K., 2002. Translating named entities using monolingual and bilingual resources. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania.

Berger, A., Della Pietra, S., Della Pietra, V., 1996. Maximum entropy approach to natural language processing. Computat. Ling. 22 (1), 39–71.

Bhagat, R., Hovy, E., 2007. Phonetic models for generating spelling variants. In: Proceedings International Joint Conference of Artificial Intelligence (IJCAI). Hyderabad, India.

Bhagat, R., Ravichandran, D., 2008. Large scale acquisition of paraphrases for learning surface patterns. In: Proceedings of ACL 2008. pp. 674–682.

Callison-Burch, C., 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In: Proceedings of EMNLP.

Callison-Burch, C., Koehn, P., Osborne, M., 2006. Improved statistical machine translation using paraphrases. In: Proceedings of NAACL-2006.

Chan, T.P., Callison-Burch, C., Van Durme, B., 2011. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics. pp. 33–42.

Chen, S., Rosenfeld, R., 2000. A survey of smoothing techniques for ME models. IEEE Trans. Speech Audio Process.

Das, D., Petrov, S., 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA. pp. 600–609. http://www.aclweb.org/anthology/P11-1061.

Denkowski, M., Al-Haj, H., Lavie, A., 2010. Turker-assisted paraphrasing for English–Arabic machine translation. In: Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 66–70.

Firth, J.R., 1957. A synopsis of linguistic theory 1930–1955. In: Studies in Linguistic Analysis. Philological Society, Oxford, pp. 1–32 (Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952–1959, London: Longman (1968)).

Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., Roukos. S., 2004. A statistical model for multilingual entity detection and tracking. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. pp. 1–8.

Ge, N., 2004. Improvements in word alignments. Presentation given at DARPA/TIDES NIST MT Evaluation workshop.

Goldsmith, J., 2001. Unsupervised learning of the morphology and natural language. Computational Linguistics 27 (2), 153–198.

Goodman. J., 2002. Sequential conditional generalized iterative scaling. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania.

Harris, Z., 1954. Distributional structure. Word 10 (23), 146–162.

Li, H., Zheng, J., Ji, H., Li, Q., Wang, W., 2013. Name-aware machine translation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria. pp. 604–614.

Huang, F., Emami, A., Zitouni, I., 2008. When Harry met Harri: Cross-lingual name spelling normalization, EMNLP'08. October 25–27, Waikiki, Honolulu, Hawaii.

Huang, F., Vogel, S., Waibel, A., 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In: Proceedings of the ACL 2003 Workshop on Multilingual and Mixed Language Named Entity Recognition – Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ.

Ide, N., Erjavec, T., Tufis, D., 2002. Sense discrimination with parallel corpora. In: Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. pp. 61–66.

Izwaini, S. (ed.), 2011. Romanization of Arabic names. Abu Dhabi: UAE Ministry of Culture, Youth and Community Development. ISBN 978-9948-15-988-9.

Ji, H., Grishman. R., 2007. Collaborative entity extraction and translation. In: Proc. International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria.

Knuth, D., 1973. The Art of Computer Programming – Volume 3: Sorting and Searching. Addison-Wesley Publishing Company.

Lee, Y.-S., Papineni, K., Roukos, S., Emam, O., Hassan, H., 2003. Language model based Arabic word segmentation. In: Proceedings of the ACL'03. pp. 399–406.

Linden, K., 2006. Multilingual modeling of cross-lingual spelling variants. Inf. Retrieval 9 (3), 295–310.

Madnani, N., Ayan, N.F, Resnik, P., Dorr, B., 2007. Using paraphrases for parameter tuning in statistical machine translation. In: Proc. WMT.

Madnani, N., Resnik, P., Dorr, B., Schwartz, R., 2008. Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. In: Proc. AMTA.

Madnani, N., Dorr, B., 2013. Generating targeted paraphrases for improved translation. ACM Trans. Intell. Syst. Technol. 4 (3).

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Marton, Y., Callison-Burch, C., Resnik, P, 2009. Improved statistical machine translation using monolingually-derived paraphrases. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore.

Marton, Y., 2010. Improved statistical machine translation using monolingual text and a shallow lexical resource for hybrid phrasal paraphrase generation. In: The Ninth Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado.

Marton, Y., 2013. Distributional phrasal paraphrase generation for statistical machine translation. In: Haifeng Wang, Bill Dolan, Idan Szpektor, Shiqi Zhao, (eds.), ACM Trans. Intell. Syst. Technol. (TIST) special issue on paraphrasing 4(3).

McCallum, A., Nigram, K., 1998. Employing EM in pool-based active learning for text classification. In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization.

NIST, 2007. The ACE Evaluation Plan. www.nist.gov/speech/tests/ace/index.htm.

Noreen, E.W., 1989. Computer-Intensive Methods for Testing Hypothesis. John Wiley Sons.

Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J., 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109–022), IBM Research Division, Thomas J. Watson Research Center.

Pasca, M., Dienes, P., 2005. Aligning needles in a Haystack: paraphrase acquisition across the web. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), Jeju Island, Republic of Korea, pp. 119–130.

Raghavan, H., Allan, J., 2005. Matching inconsistently spelled names in automatic speech recognizer output for information retrieval. In: Proceedings of HLT/EMNLP. pp. 451–458.

Ramshaw, L., Marcus, M., 1994. Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. In: Proceedings of the ACL Workshop on Combining Symbolic and Statistical Approaches to Language. pp. 128–135.

Rogati, M., McCarley, S., Yang, Y., 2003. Unsupervised learning of Arabic Stemming using a parallel corpus. In: Proceedings of ACL'03, Sapporo, Japan, pp. 391–398.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas.

Tackstromand, O., McDonald, R., Uszkoreit, J., 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).

Tjong Kim Sang, E.F., 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158.

Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., Järvelin, K., 2005. Translating cross-lingual spelling variants using transformation rules. Inf. Process. Manage. 41 (4), 859–872.

Vogel, S., Ney, H., Tillmann, C., 1996. HMM-based word alignment in statistical translation. In: Proceedings of the 16th Conference on Computational Linguistics – Volume 2 (Copenhagen, Denmark, August 05–09, 1996). International Conference on Computational Linguistics. Morristown, NJ.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised models. In: Proceedings of ACL'95.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., Callison-Burch, C., 2012. Machine translation of Arabic dialects. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 49–59.

Zitouni, I., Florian R., 2008. Mention detection crossing the language barrier. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Waikiki, Honolulu, Hawaii.

Zitouni, I., Sorensen, J., Luo, X., Florian, R., 2005. The impact of morphological stemming on Arabic mention detection and coreference resolution. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. The 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor.

Please cite this article in press as: Marton, Y., Zitouni, I. Transliteration normalization for Information Extraction and Machine Translation. Journal of King Saud University – Computer and Information Sciences (2014), http://dx.doi.org/10.1016/j.jksuci.2014.06.011