RC24739 (W0902-038) February 5, 2009
Computer Science

IBM Research Report

When Harry Met Harri, هاري and 亨利 : Cross-lingual Name Spelling Normalization

Fei Huang, Ahmad Emami, Imed Zitouni
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

When Harry Met Harri, هاري and 亨利: Cross-lingual Name Spelling Normalization

Fei Huang , Ahmad Emami and Imed Zitouni

IBM T. J. Watson Research Center

1101 Kitchawan Road
Yorktown Heights, NY 10598
{huangfe, emami, izitouni}@us.ibm.com

Abstract

Foreign name translations typically include multiple spelling variants. These variants cause data sparseness problems, increase Out-of-Vocabulary (OOV) rate, and present challenges for machine translation, information extraction and other NLP tasks. This paper aims to identify name spelling variants in the target language using the source name as an anchor. Based on word-to-word translation and transliteration probabilities, as well as the string edit distance metric, target name translations with similar spellings are clustered. With this approach tens of thousands of high precision name translation spelling variants are extracted from sentence-aligned bilingual corpora. When these name spelling variants are applied to Machine Translation and Information Extraction tasks, improvements over strong baseline systems are observed in both cases.

1 Introduction

Foreign names typically have multiple spelling variants after translation, as seen in the following examples:

He confirmed that "***al-Kharroub*** province is at the top of our priorities."

…for the Socialist Progressive Party in upper Shuf and the ***Al-Kharrub*** region,…

…during his tour of a number of villages in the region of ***Al-Kharub***,…

…Beirut and its suburbs and Iqlim al-Khurub,…

Such name spelling variants also frequently appear in other languages, such as 布什(bushi) / 布殊(bushu) / 布希(buxi) (for Bush) in Chinese, and سبرنغفيلد (sbrngfyld) /سبرينغفيلد (sbryngfyld) / سبرينجفيلد (sbrynjfyld) (for Springfield) in Arabic.

These spelling variants present challenges for many NLP tasks, increasing vocabulary size and OOV rate, exacerbating the data sparseness problem and reducing the readability of MT output when different spelling variants are generated for the same name in one document. We address this problem by replacing each spelling variant with its corresponding canonical form. Such text normalization could potentially benefit many NLP tasks including information retrieval, information extraction, question answering, speech recognition and machine translation.

Research on name spelling variants has been studied mostly in Information Retrieval research, especially in query expansion and cross-lingual IR. Baghat and Hovy (2007) proposed two approaches for spelling variants generation, based on the letters-to-phonemes mapping and Soundex algorithm (Knuth 1973). Raghaven and Allan (2005) proposed several techniques to group names in 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 (2005) proposed to use 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 (2003) and Ji and Grishman (2007) investigated the general name entity translation problem, especially in the context of machine translation.

This paper aims to identify mono-lingual name spelling variants using cross-lingual information. Instead of using a named entity tagger to identify name spelling variants, we treat names in one language as the anchor of spelling variants in another language. From sentence-aligned bilingual corpora we collect word co-occurrence statistics and calculate word translation[1] probabilities. For each source word, we group its target translations into clusters according to string edit distances, then calculate the transliteration cost between the source word and each target translation cluster. Word pairs with small transliteration costs are considered as name translations, and the target cluster contains multiple spelling variants corresponding to the source name.

We apply this approach to extract name transliteration spelling variants from bilingual corpora. We obtained tens of thousands of high precision name translation pairs. We further apply these spelling variants to Machine Translation (MT) and Information Extraction (IE) tasks, and observed statistically significant improvement on the IE task, and close to oracle improvement on the MT task.

The rest of the paper is organized as follows. In section 2 we describe the technique to identify name spelling variants from bilingual data. In section 3 and 4 we address their application to MT and IE respectively. We present our experiment results and detailed analysis in section 5. Section 6 concludes this paper with future work.

2 Finding Name Translation Variants

[1] In this paper, the translation cost measures the semantic difference between source and target names, which are estimated from their co-occurrence statistics. The transliteration cost measures their phonetic distance and are estimated based on a character transliteration model.

Starting from sentence-aligned parallel data, we run HMM alignment (Vogel et. al. 1996 & Ge 2004) to obtain a word translation model. For each source word this model generates target candidate translations as well as their translation probabilities. A typical entry is shown in Table 1. It can be observed that the Arabic name's translations include several English words with similar spellings, all of which are correct translations. However, because the lexical translation probabilities are distributed among these variants, none of them has the highest probability. As a result, the incorrect translation, *iqlim*, is assigned the highest probability and often selected in MT output. To fix this problem, it is desirable to identify and group these target spelling variants, convert them into a canonical form and merge their 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]

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

For each source word in the word translation model, we cluster its target translations based on string edit distances using group average agglomerative clustering algorithm (Manning and Schütze, 2000). Initially each target word is a single word 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 threshold. In the above example, *al-kharrub, al-kharub, al-khurub* and *al-kharroub* are grouped into a single cluster, and each of the ungrouped words remains in its 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 اعلم which is aligned to 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 transliteration cost between the source word and each target word in the cluster. As

many names are translated based on their pronunciations, the source and target names have similar phonetic features and lower transliteration costs. Word pairs whose transliteration cost is lower than an empirically selected threshold are considered as name translations.

2.1 Name Transliteration Cost

The transliteration cost measures the phonetic similarity between a source word and a target word. It is calculated based on the character transliteration model, which can be trained from bilingual name translation pairs. We segment the source and target names into characters, then run monotone[2] HMM alignment on the source and target character pairs. 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,…, n, letter e_j is aligned to character f_{a_j} according to the HMM aligner. Under the assumption that character alignments are independent, the word transliteration probability is calculated as

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

where $p(e_j \mid f_{a_j})$ 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.

An example of the trained A-E character transliteration model is shown in Figure 1. The Arabic character ż is aligned with high probabilities to English letters with similar pronunciation. Because Arabic words typically omit vowels, English vowels are also aligned to Arabic characters. Given this model, the characters within a Romanized Arabic name and its English translation are aligned as shown in Figure 1.

[2] As name are typically phonetically translated, the character alignment are often monotone. There is no cross-link in character alignments.

2.2 Transliteration Unit Selection

The transliteration units are typically characters. The Arabic alphabet includes 32 characters, and the English alphbet includes 56 letters [3]. However, Chinese has about 4000 frequent characters. The imbalance of Chinese and English vocabulary sizes results in suboptimal transliteration model estimation. Each Chinese character also has a pinyin, the Romanized representation of its pronunciation. Segmenting the Chinese pinyin into sequence of Roman letters, we now have comparable vocabulary sizes for both Chinese and English. We build a pinyin transliteration model using Chinese-English name translation pairs, and compare its performance with a character transliteration model in Experiment section 5.1.

ż | x

h	K	k	a	u	i
[0.44]	[0.29]	[0.21]	[0.03]	[0.015]	[0.004]

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

3 Application to Machine Translation

We applied the extracted name translation spelling variants to the machine translation task. Given the name spelling variants, we updated both the translation and the language model, adding variants' probabilities to the canonical form.

Our baseline MT decoder is a phrase-based decoder as described 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 translation model cost, language model cost, distortion cost and

[3] Uppercase and lowercase letters plus some special symbols such as '_', '-".

sentence length cost. The translation cost includes word translation probability and phrase translation probability.

3.1 Updating The Translation Model

Given target name spelling variants $\{t_1, t_2, ..., t_m\}$ for a source name *s*, here $t_1, t_2, ..., t_m$ are sorted based on their lexical translation probabilities, $p(t_1 \mid s) \geq p(t_2 \mid s) \geq ... \geq p(t_m \mid s)$.

We select t_1 as the canonical spelling, and merge other spellings' translation probabilities with this one:

$$p(t_1 \mid s) = \sum_{j=1}^{m} p(t_m \mid s).$$

Other spelling variants get zero probability. Table 2 shows the updated word translation probabilities for "الخروب|Alxwrb". Compared with Figure 1, the translation probabilities from several spelling variants are merged with the canonical form, *al-kharrub*, which now has the highest probability in the new model.

الخروب | Alxwrb

al-kharrub [0.35]	iqlim [0.22]	al-kharub [0.0]	overflew [0.09]
junbulat [0.05]	al-khurub [0.0]	hours [0.04]	al-kharroub [0.0]

Table 2. English translations of an Arabic name الخروب|*Alxrwb* with the updated word translation model.

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 source name and its spelling variant in the target phrase, we replace the target name with its canonical spelling. After the mapping, two target phrases differing only in target names may end up with the identical target phrase, and their alignment frequencies are added. Phrase translation probabilities are re-estimated with the updated frequencies.

3.2 Updating The Language Model

The machine translation decoder uses a language model as a measure of a well-formedness of the output sentence. Since the updated translation model can produce only the canonical form of a group of spelling variants, the language model should be updated in that all *m*-grams ($1 \leq m \leq N$) that are spelling variants of each other are merged (and their counts added), resulting in the canonical form of the *m*-gram. 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 that t_1^i and t_2^i belong to the same spelling variant group.

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 build the language model again. However, since 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 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.

4 Applying 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*[4]: we call instances of textual references to objects *mentions*, which can be either named (e.g. John Smith), nominal (the president) or pronominal (e.g. he, she). For instance, in the sentence

- President *John Smith* said *he* has no comments.

there are two mentions: *John Smith* and *he*. Similar to many classical NLP tasks, we formulate the mention detection problem as a classification problem, by assigning to each token in the text a label, indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. Good

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

performance in many natural language processing tasks has been shown to depend heavily on integrating many sources of information (Florian et al. 2007). We select an exponential classifier, the Maximum Entropy (MaxEnt henceforth) classifier that 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, syntactic, and the output of other information extraction models. These features were described in (Zitouni and Florian, 2008 & Florian et al. 2007), and are not discussed here. In this paper we focus on examining the effectiveness of name spelling variants in improving mention detection systems. We add a new feature that for each token x_i to process we fire its canonical form (class label) $C(x_i)$, representative of name spelling variants of x_i. This name spelling variant feature is also used in *conjunction* with the lexical (e.g., words and morphs in a 3-word window, prefixes and suffixes of length up to 4, stems in a 4-word window for Arabic) and syntactic (POS tags, text chunks) features.

5 Experiments

5.1 Evaluating the precision of name spelling variants

We extracted Arabic-English and English-Arabic name translation variants from sentence-aligned parallel corpora released by LDC. The accuracy of the extracted name translation spelling variants are judged by proficient Arabic and Chinese speakers.

The Arabic-English parallel corpora include 5.6M sentence pairs, 845K unique Arabic words and 403K unique English words. We trained a word translation model by running HMM alignment on the parallel data, grouped target translation with similar spellings and computed 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 position for evaluation (20 samples at each ranking position). The quality of the name translation variants are judged as follows: for each candidate name translation group $\{t_1, t_2, ..., t_m \mid s\}$, if the source word *s* is a name and all the target spelling variants are correct translations, it gets 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 gets partial credit *n/m*, where *n* is the number of correct target translations. We evaluate only the precision of the extracted spelling variants[5]. As seen in Figure 2, the precision of the top 22K A-E name translations is 96.9%. Among them 98.5% of the Arabic words are names. The precision gets lower and lower when 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.

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 Figure 3, top 20K E-A 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 A-E and E-A name spelling variants, where Arabic words are represented in their Romanized form.

We conduct a similar experiment on the Chinese-English language pair, extracting Chinese-English and English-Chinese name spelling variants from 8.7M Chinese-English sentence pairs. After word segmentation, the Chinese vocabulary size is 1.5M words, and English vocabulary size is 1.4M words. With the

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

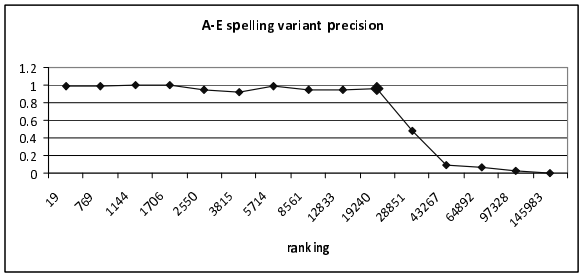

Figure 2. Arabic-English name spelling variants precision curve (Precision of evaluation sample at different ranking positions. The larger square indicates the cutoff point).

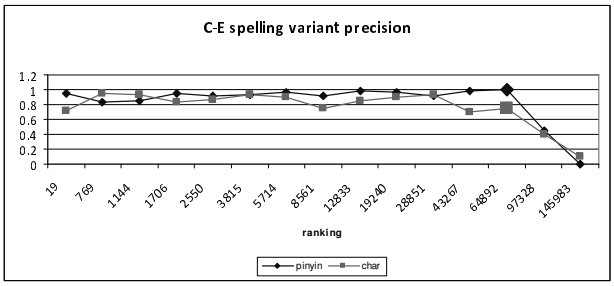

Figure 4. Chinese-English name spelling variants precision curve.

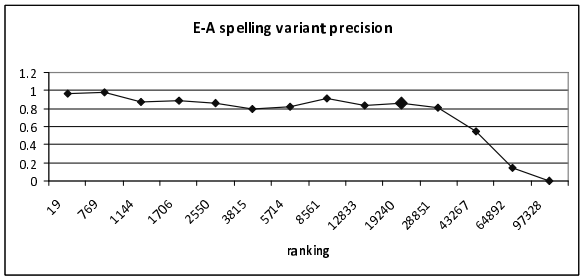

Figure 3. English-Arabic name spelling variants precision curve.

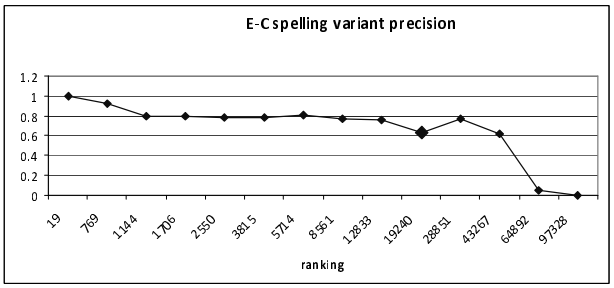

Figure 5. English-Chinese name spelling variants precision curve.

Chinese pinyin transliteration model, we extract 64K C-E name spelling variants with 93.6% precision. Figure 4 also shows the precision curve of the Chinese character transliteration model. On average the pinyin transliteration model has about 6% higher precision than the character transliteration model. The pinyin transliteration model is particularly better on the tail of the curve, extracting more C-E transliteration variants. Figure 5 shows the precision curve for E-C name spelling variants, where 20K name pairs are extracted using letter-to-character transliteration model, and obtaining a precision of 74.3%.

Table 4 shows some C-E and E-C name spelling variants. We observed errors due to word segmentation. For example, the last two Chinese words corresponding to "drenica" have additional Chinese characters, meaning "drenica region" and "drenica river". Similarly for tenet, the last two Chinese words also have segmentation errors due to missing or spurious characters. Note that in the C-E spelling variants, the source word "韦尔曼" has 14 spelling variants. Judge solely from the spelling, it is hard to tell whether they are the same person name with different spellings.

5.2 Experiments on Machine Translation

We apply the Arabic-English name spelling variants on the machine translation task. Our baseline system is trained with 5.6M Arabic-English sentence pairs, the same training data used to extract A-E spelling variants. The language model is a modified Kneser-Ney 5-gram model trained on roughly 3.5 billion words. After pruning (using count cutoffs), it contains a total of 935 million *N*-grams. We updated the translation models and the language model with the name spelling variant class.

Table 5 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 translation and language models assign higher probabilities to the correct translation "al-kharroub" after spelling variant normalization.

Lang. Pair	Source Name	Target Spelling Variants
Arabic-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-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

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

Lang. Pair	Source Name	Target Spelling Variants
Chinese-English	延多维茨基 (yan/duo/wei/ci/ji)	endovitsky jendovitski yendovitski endovitski
	斯特凡尼 (si/te/fan/ni)	stefani steffani stephani stefanni stefania
	韦尔曼 (wei/er/man)	woermann wellman welman woellmann wohrmann wormann velman wollmann wehrmann verman woehrmann wellmann welmann wermann
English-Chinese	tenet	特尼特(te/ni/te) 特内特(te/nei/te) 泰内特(tai/nei/te) 特耐特(te/nai/te) 特奈特(te/nai/te) *特内特于(te/nei/te/yu) 特内(te/nei)*
	drenica	德雷尼察(de/lei/ni/cha) 德雷尼卡(de/lei/ni/ka) 特雷尼察(te/lei/ni/cha) 特雷尼查(te/lei/ni/cha) *德雷尼察区(de/lei/ni/cha/qu) 德雷尼察河 (de/lei/ni/cha/he)*
	ahmedabad	艾哈迈达巴德(ai/ha/mai/da/ba/de) 艾阿迈达巴德(ai/a/mai/da/ba/de) 艾哈默德巴德(ai/ha/mo/de/ba/de) 阿哈迈达巴德(a/ha/mai/da/ba/de)

Table 4. Chinese-English and English-Chinese name spelling variant examples with pinyin for Chinese characters. *Italic words* represent errors due to word segmentation.

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

Table 5. English translation output with the baseline MT system and the system with updated models

	BLEU r1n4	TER
Baseline	0.2714	51.66
Baseline+ULM+UTM	0.2718	51.46
Ref. Normalization	0.2724	51.40

Table 6. MT scores with updated TM and LM

We also evaluated the updated MT models on a MT test set. The test set includes 70 documents selected from GALE 2007 Development set. It contains 42 newswire documents and 28 weblog and newsgroup documents. There are 669 sentences with 16.3K Arabic words in the test data. MT results are 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 6. Applying the updated language model (ULM) and the translation model (UTM) lead to a small reduction in TER. After we apply similar name spelling normalization on the reference translation, we observed some additional improvements. Overall, the BLEU score is increased by 0.1 BLEU point and TER is reduced by 0.26.

Although the significance of correct name translation can not be fully represented by

BLEU and TER scores [6], we still want to understand the reason of the relatively small improvement. After some error analysis, we found that in the testset only 2.5% of 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. Because the MT system is trained on the same bilingual data from which the name spelling variants are extracted, some of these Arabic names are already correctly translated in the baseline system. So the room of improvement is small. We did an oracle experiment, manually correcting the name translation errors in the first 10 documents (89 sentences with 2545 words). With only 6 name translation errors corrected, this reduced the TER from 48.83 to 48.65.

5.2 Experiments on Information Extraction

Mention detection system experiments are conducted on the ACE 2007 data sets in Arabic and English. Since the evaluation test set is not publicly available, we have split the publicly available training corpus into an 85%/15% data split. To facilitate future comparisons with work presented here, and to simulate a realistic scenario, the splits are created based on article dates: the test data is selected as the latest 15% of the data in chronological order. 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 are using a large range of features, including lexical (e.g., words and morphs in a 3-word window, prefixes and suffixes of length up to 4, stems in a 4-word window for Arabic), syntactic (POS tags, text chunks), and the output of other information extraction models. These features were described in (Zitouni and Florian, 2008 & Florian et al. 2007) with more details. Our goal here is to investigate the effectiveness of name spelling variants information in improving mention detection system performance.

	Baseline			**Baseline+NSV**		
	P	**R**	**F**	**P**	**R**	**F**
English	84.4	80.6	**82.4**	84.6	80.9	**82.7**
Arabic	84.3	79.0	**81.6**	84.4	79.1	**81.7**

Table 7: Performance of English and Arabic mention detection systems without (Baseline) and with (Baseline+NSV) the use of name spelling variants. Performance is presented in terms of Precision (P), Recall (R), and F-measure (F).

Results in Table 7 show that the use of name spelling variants (**NSV**) improves mention detection systems performance, especially for English; an interesting improvement is obtained in recall – which is to be expected, given the method –, but also in precision, leading to systems with better performance in terms of F-measure (82.4 vs. 82.7). This improvement in performance is statistically significant according to the stratified bootstrap re-sampling approach (Noreen 1989). This approach is used in the named entity recognition shared task of CoNLL-2002 [7]. However, the small improvement obtained for Arabic is not statistically significant based on the approach described earlier. One hypothesis 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 Conclusion

We proposed a cross-lingual name spelling variants extraction technique. We extracted tens of thousands of high precision bilingual name translation spelling variants. We applied the spelling variants to the IE task, observing statistically significant improvements over a strong baseline system. We also applied the spelling variants to MT task and even though the overall improvement is relatively small, it achieves performance close to the one observed in an oracle experiment.

[6] These scores treat information bearing words, like names, the same as any other words, like punctuations.

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

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-2-0001 under the GALE program. We are grateful to Yaser Al-Onaizan, Salim Roukos and anonymous reviewers for their constructive comments.

References

Al-Onaizan, Y. and Papineni, K. Distortion Models for Statistical Machine Translation. In Proceedings of the 44th Annual Meeting on Association For Computational Linguistics. Sydney, Australia. July 2006.

Al-Onaizan, Y. and Knight, K. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics* (Philadelphia, Pennsylvania, July 07 - 12, 2002). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ. 2002

Berger, A., S. Della Pietra, and V. Della Pietra. A Maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71. 1996

Bhagat, R. and Hovy, E. "Phonetic Models for Generating Spelling Variants", *In Proceedings International Joint Conference of Artificial Intelligence* (IJCAI). Hyderabad, India. 2007.

Chen, S. and Rosenfeld R. A survey of smoothing techniquesfor ME models. IEEE Trans. On Speech and Audio Processing. 2002

Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., and Roukos. S. 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, pages 1–8.

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

Goodman. J. Sequential conditional generalized iterative scaling. In *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics* (Philadelphia, Pennsylvania, July 07 - 12, 2002). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ. 2002

Huang, F., Vogel, S., and Waibel, A. 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, 2003

Ji, H. and Grishman. R. Collaborative Entity Extraction and Translation. *Proc. International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria. Sept 2007.

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

Linden, K. "Multilingual Modeling of Cross-Lingual Spelling Variants", *Information Retrieval*, Vol. 9, No. 3. (June 2006), pp. 295-310.

Manning, C.D., and Schutze., H. *Foundations of Statistical Natural Language Processing*. MIT Press, 2000

NIST. 2007. The ACE evaluation plan. www.nist.gov/speech/tests/ace/index.htm.

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

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

Raghavan, H. and Allan, J., "Matching Inconsistently Spelled Names in Automatic Speech Recognizer Output for Information Retrieval," *the Proceedings of HLT/EMNLP* 2005, pp. 451-458.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. "A Study of Translation Edit Rate with Targeted Human Annotation," *Proceedings of Association for Machine Translation in the Americas*, 2006.

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

Vogel, S., Ney, H., and Tillmann, C.. 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. Association for Computational Linguistics, Morristown, NJ 1996

Zitouni, I., Florian R.. Mention Detection Crossing the Language Barrier. Proceedings of Conference on Empirical Methods in Natural Language Processing. Waikiki, Honolulu, Hawaii (October, 2008)

Zitouni, I., Sorensen, J., Luo, X., and Florian, R. 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 (June, 2005)