A General Method for Creating a Bilingual Transliteration Dictionary

Amit Kirschenbaum, Shuly Wintner

Department of Computer Science, Department of Computer Science
University of Leipzig, University of Haifa
amit@informatik.uni-leipzig.de, shuly@cs.haifa.ac.il

Abstract

Transliteration is the rendering in one language of terms from another language (and, possibly, another writing system), approximating spelling and/or phonetic equivalents between the two languages. A transliteration dictionary is a crucial resource for a variety of natural language applications, most notably machine translation. We describe a general method for creating bilingual transliteration dictionaries from Wikipedia article titles. The method can be applied to any language pair with Wikipedia presence, independently of the writing systems involved, and requires only a single simple resource that can be provided by any literate bilingual speaker. It was successfully applied to extract a Hebrew-English transliteration dictionary which, when incorporated in a machine translation system, indeed improved its performance.

1. Introduction

Transliteration is the process of converting terms written in one language into their approximate spelling or phonetic equivalents in another language. Transliteration is defined for a pair of languages, a *source* language and a *target* language. The two languages may differ in their writing systems and phonetic inventories.

Transliteration has acquired a growing interest recently, particularly in the field of Machine Translation (MT) (Stalls and Knight, 1998; Al-Onaizan and Knight, 2002; Yoon et al., 2007; Hermjakob et al., 2008). It handles those terms where no translation would suffice or even exist. Failing to recognize such terms would result in poor performance of the translation system.

Bilingual transliteration dictionaries are therefore essential resources for various natural language processing applications, notably MT. They are also useful as training corpora for machine-learning-based applications that induce grapheme sequence correspondences across languages, such as MT and named entity recognition (Goldwasser and Roth, 2008; Kirschenbaum and Wintner, 2009; Li and Kumaran, 2009). Construction of such dictionaries is naturally time-consuming, labor-intensive and expensive.

We describe in Section 2. a general method for creating bilingual transliteration dictionaries from Wikipedia article titles. The method can be applied to any two languages represented in Wikipedia, independently of the writing systems involved.[1] It requires only a single simple resource, a consonant-correspondence table, that can be quickly and easily constructed. The method was successfully applied to extract a Hebrew-English transliteration dictionary (Section 3.) which, when incorporated in a machine translation system, indeed improved its performance (Section 4.). The dictionary, which is publicly-available[2], will also be used for a transliteration shared task, as part of the 2010 NEWS workshop.

[1] We do conjecture, however, that for languages whose writing systems use syllabaries or logograms our method is sub-optimal.

[2] http://cl.haifa.ac.il/projects/mt/

2. Methodology

The method we describe here is general, and can be applied to any language pair with significance Wikipedia presence. Currently, 31 languages are represented with over 100,000 documents on Wikipedia; 92 with over 10,000 documents; 186 with over 1,000 documents, and 246 with over 100 documents.[3] We estimate that as few as a couple of thousands of parallel documents are sufficient to generate a useful transliteration dictionary (in our experience, the number of transliteration pairs extracted from the corpus is approximately half the number of article pairs).

We extract a parallel list of source- and target-language terms from Wikipedia to automatically generate a corpus of transliteration terms. Wikipedia documents in several languages are linked to one another by explicit links, so the meta-information provided in the documents is sufficient to determine that two documents discuss the same topic. However, the documents are not necessarily translations of one another (they should be considered *comparable*, rather than *parallel*). We therefore do not use the contents of the documents to extract term-pairs; rather, we focus only on the *title* of the documents. These are necessarily co-references to the same entity or concept in the two languages, and are therefore with high likelihood either translations or transliterations of each other.

In order to determine whether title terms are translations or transliterations, we employ a measure of similarity of consonants in parallel source and target terms. The similarity measure is based only on consonants since vowel correspondences across languages tend to be less predictable; specifically, vowels are often not represented at all in Hebrew (and other languages, e.g., Arabic and Persian).[4]

To facilitate readability, we depict Hebrew letters in this paper using ASCII characters through the 1-1 mapping listed in the following table:

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias, retrieved 2010/02/25.

[4] Languages whose writing systems use syllabaries or logograms may require more sophisticated correspondences; we ignore this complication here.

א	ב	ג	ד	ה	ו	ז	ח	ט	י	כ
a	b	g	d	h	w	z	x	@	i	k
ל	מ	נ	ס	ע	פ	צ	ק	ר	ש	ת
l	m	n	s	&	p	c	q	r	$	t

To determine consonant correspondences between Hebrew and English we constructed a simple table, based on common knowledge patterns that relate sound to spelling in both languages. Sound patterns that are not part of the phoneme inventory of Hebrew but are nonetheless represented in Hebrew orthography (g', z' and c') were also included in the table. Every entry in the mapping table consists of a Hebrew letter and a possible English letter or letter sequence that might match it. A typical entry is the following:

K:K|CH|KH

implying that the Hebrew consonant 'k' can be matched to 'k', 'ch' or 'kh' in English. Crucially, such correspondence tables can be built for any language pair, by any literate bilingual speaker, in a matter of minutes. The table we constructed for Hebrew-English, which consists of 33 entries, is listed in Table 1.

B:B|V|W	0:0
G':G|J	1:1
G:G	2:2
D:D|T	3:3
H:H	4:4
W:V|W	5:5
Z':G|J|ZH	6:6
Z:Z|S	7:7
X:CH|H|KH	8:8
@:T	9:9
KS:X|KS	
K:K|CH|KH	
L:L	
M:M	
N:N	
S:S|C	
P:P|F|V	
C':CH|CZ	
C:TS|TZ|C|T	
QS:X|KS	
Q:C|K|Q	
R:R	
$:SH|CH|SCH|S|C	
T:TH|T	

Table 1: Consonant correspondences between Hebrew (on the left of the ':') and English

Given two candidate terms, w_h in Hebrew and w_e in English, we iterate over the consonants of w_h (as defined by Table 1), and for each Hebrew consonant c_h we look for a matching consonant c_e in the current position of w_e, again skipping over vowels, where matching is defined by Table 1. We tally the number of such correspondences, and if the ratio of the number of consonant correspondences to the number of consonants in the Hebrew term is greater than $3/4$, we determine that the term pair is indeed a transliteration pair. The ratio was chosen empirically: a higher ratio would guarantee better precision, whereas a lower one would improve recall.

Wikipedia titles are short, but both source and target titles may consist of several words (in our case, 28,096 out of 41,913, or 67% of the entries, were multi-word). Furthermore, words composing the entries in each of the languages may be ordered differently. Therefore, every word in the source language is compared with every word in the target, assuming that titles are short enough.[5]

The example in Table 2 presents an aligned pair of multilingual Wikipedia entries with high (and monotonous) similarity of consonants. It is therefore considered as a transliteration pair.

g	r	a		t	e	f	u	l		d	e	a	d	
g	r	i	i	@		p	w	l		d		d		

Table 2: Titles of Wikipedia entries

In contrast, the title *empty set*, which is translated to *hqbwch hriqh*, shows a low similarity of consonants. This pair is not extracted as a transliteration instance.

3. Results

The 41913 Hebrew and English terms retrieved from Wikipedia yielded 19,901 that were determined as transliteration pairs; some of those are multi-word, and the total number of transliteration token pairs was 20,184 (approximately half of the number of document pairs in the corpus). It is important to note that while many of the extracted pairs are named entities, many others are not. These include, for example, Hebrew terms that are used also in English (e.g., *kaddish*) or loan words that Hebrew borrowed from other languages (e.g., *pirwmanih* "pyromania"). Furthermore, while the vast majority of entries are originally either Hebrew or English, many originate in several other languages, and some use characters that are not in the English alphabet, including diacritics etc. Figure 1 depicts some of the obtained results; the leftmost column depicts the original Hebrew form, the middle column is the English transliteration and the rightmost column depicts Hebrew using ASCII.

A robust evaluation of the quality of the transliteration pairs retrieved using this methodology is difficult, mainly because for many of the terms, several transliterated forms are valid. Nevertheless, we selected a sample of 1000 pairs, and manually verified that they were valid transliterations. Of the 1000, only 32 were poor transliterations, and approximately half of those were strings consisting mostly or solely of numbers. Such strings can of course be removed automatically. We conjecture that setting the threshold at $3/4$ matched consonants, as we did, greatly improved the precision of the method, possibly at the expense of lower recall.

[5] Incidentally, the longest title in our corpus is "French Revolution from the abolition of feudalism to the Civil Constitution of the Clergy", a 14-token string, but this is exceptional. We simply ignore entries longer than three tokens.

שטפאנ	stefan	$@pan
קופה	copa	qwph
אסמבלאז'	assemblage	asmblaz'
גליפטודונ	glyptodon	glip@wdwn
ברא	bara	bra
שרינג	szeryng	$ring
טונקס	tonks	@wnqs
שלזינגר	schlesinger	$lzingr
כיסופימ	kissufim	kiswpim
קלופט	klüft	qlwp@
ניקולס	nicholls	niqwls
אקוומרינ	aquamarine	aqwwmrin
חוסייני	husseini	xwsiini
פפטידימ	peptide	pp@idim
מגנטו	magneto	mgn@w
קרומ	krum	qrwm
דאפי	duffy	dapi
פוקס	fawkes	pwqs
פרנתרופוס	paranthropus	prntrwpws
בוגנוויליה	bougainvillea	bwgnwwilih
ראדו	radu	radw
סיליקט	silicate	siliq@
דוכובני	duchovny	dwkwbni
קיסינג'ר	kissinger	qising'r
אליהו	elio	alihw
לברדוריט	labradorite	lbrdwri@
צ'דוויק	chadwick	c'dwwiq
אנציקלופדיה	encyclopædia	anciqlwpdih
גורה	góra	gwrh
ניוקומנ	newcomen	niwqwmn
יוגורט	yogurt	iwgwr@
אפי	effi	api
נואל	noël	nwal
מונטאנ	montand	mwn@an
אפאזיה	aphasia	apazih
קרקס	circus	qrqs
וסטהיימר	westheimer	ws@hiimr

Figure 1: Results: example transliterated pairs

4. Applications

The transliteration dictionary that we described above was used as an additional dictionary that was incorporated in a Hebrew to English statistical machine-translation system (Lavie et al., 2004b). To evaluate its contribution, we ran the system with and without the dictionary. The experimental setup is a transfer-based SMT system whose parameters are tuned on a set of 500 sentences, and which is evaluated on a set of 300 sentences, each with four reference translations. We report both Meteor scores (Lavie et al., 2004a), and BLEU scores (Papineni et al., 2002). The results, which are depicted in Table 3, show a statistically significant improvement in translation accuracy ($p < 0.05$).

The dictionary was also used as a corpus for training a machine-learning-based Hebrew to English transliteration system (Kirschenbaum and Wintner, 2009). The transliteration system was then added as a module to the Hebrew to English statistical machine-translation system. Source-language tokens are fed to a classifier which determines whether they should be translated or transliterated; in the latter case, a transliteration module trained on the dictionary described above is used to generate up to 10 transliteration candidates which are fed to the translation engine. Again, the quality of the translations (evaluated using BLEU and Meteor scores) improves significantly when the transliteration module is added.

System	BLEU	METEOR
Baseline	15.21	36.94
With transliteration dictionary	15.54	37.23

Table 3: Improvement of translation accuracy as as result of using the transliteration dictionary

In addition, the transliteration corpus we extracted will be used for the shared task on Machine Transliteration of named entities, which will be part of the Named Entities WorkShop (NEWS) planned for 2010. In principle, corpora for any language pair included in Wikipedia can be automatically generated in the same way.

5. Conclusion

We presented a general, language-independent method for extracting transliteration pairs from Wikipedia titles. We used the method to automatically create a highly accurate Hebrew-English transliteration dictionary. This dictionary is used as training material for a transliteration module that we developed, which improves the quality of Hebrew-to-English machine translation.

This work can be extended in several ways. The consonant similarity method is rather ad-hoc, and not necessarily the best possible one. We conjecture that very simple methods, based on weighted edit distance (Levenshtein, 1965; Kruskal, 1999), can work just as well or even better. Of course, we only applied the method to a single language pair, but given that Hebrew and English are very different in almost every aspect, this seems to be a challenging choice. Still, applications to more language pairs are needed in order to establish the robustness of this method. We are currently working on such an application for Hebrew-Arabic.

Acknowledgments

We wish to thank Gennadi Lembersky for his help in integrating our work into the MT system, as well as to Erik Peterson and Alon Lavie for providing the code for extracting bilingual article titles from Wikipedia. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06); by the Israel Internet Association; by the Knowledge Center for Processing Hebrew; and by the Caesarea Rothschild Institute for Interdisciplinary Application of Computer Science at the University of Haifa.

6. References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *ACL '02: Proceedings of the 40th Annual*

Meeting on Association for Computational Linguistics, pages 400–408, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Goldwasser and Dan Roth. 2008. Active sample selection for named entity transliteration. In *Proceedings of ACL-08: HLT, Short Papers*, pages 53–56, Columbus, Ohio, June. Association for Computational Linguistics.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of ACL-08: HLT*, pages 389–397, Columbus, Ohio, June. Association for Computational Linguistics.

Amit Kirschenbaum and Shuly Wintner. 2009. Lightly supervised transliteration for machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 433–441, Athens, Greece, March. Association for Computational Linguistics.

Joseph Kruskal. 1999. An overview of sequence comparison. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. CSLI Publications, Stanford, CA. Reprint, with a foreword by John Nerbonne.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004a. The significance of recall in automatic metrics for mt evaluation. In Robert E. Frederking and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 134–143. Springer.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004b. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Haizhou Li and A Kumaran, editors. 2009. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*. Association for Computational Linguistics, Suntec, Singapore, August.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41.

Su-Youn Yoon, Kyoung-Young Kim, and Richard Sproat. 2007. Multilingual transliteration using feature based phonetic method. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 112–119, Prague, Czech Republic, June. Association for Computational Linguistics.