Transliterated Pairs Acquisition in Medical Hebrew

Raphael Cohen, Yoav Goldberg and Michael Elhadad
Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel.

In most domain-specific texts encoded in a non-Latin alphabet, many proper names, named entities and open-class lexical items are transliterated from English. We investigate some of the problems caused by the high frequency of transliterations in Hebrew in the medical domain.

The phonetic transcription of a word from a source language using a different script is called transliteration. Transliterations affect Information Extraction (IE) in two ways. First, it takes time for a transliterated word to make it into a technical lexicon, making recognition difficult. A second problem is the variability of ways a foreign word can be rendered phonetically, leading in most cases (except for very short words) to many possible spellings of the word and, therefore, making lexicon-based recognition difficult. In this paper, we present a method for automatically acquiring transliterated words and their source word in order to improve a technical lexicon, addressing both problems: spelling variants and unknown tokens.

Information Extraction tasks commonly occur in specific domains (Financial, Medical, Technology). Such domains contain technical words, multi-word expressions and proper names specific to the domain. These words in the specific corpora may have different features than the words in a general corpus. In English, state of the art medical IE methods, based on medical term detection, are lexicon based (Jain et al, 1996, Teufel and Elhadad, 2002). These methods rely on UMLS (Unified Medical Language System) a well maintained collection of over 30 medical vocabularies.

In Hebrew, such a vocabulary is not available. We use a lexicon acquired from the medical terms lists in the popular medical information site Infomed.co.il. This lexicon includes 6,010 medical terms and 1,981 medication names in Hebrew.

To examine the performance of the medical term detection task, we use a small annotated corpus of Questions/Answers from Infomed.co.il, comprising 75 documents, 11,508 words and 1,530 annotated medical terms. Term detection using the Infomed lexicon and exact match yields recall of 16% and precision of 84.7%. Using the Infomed lexicon and a Hebrew Morphological Analyzer (Adler and Elhadad 2006) (not adapted to the medical domain) allows us to search lexicon terms after the morphological analyzer segmented the terms from their prefixes. The improved segmentation yielded higher recall at 33% and higher precision at 89% on instances (31% recall and 88% precision on distinct terms). The recall results remain extremely low compared to a similar simple baseline in English text.

Preliminary error analysis suggests that a significant portion of our recall errors fall on transliterated medical terms. This stems from spelling variations and segmentation errors on the unknown words. Loose matching to the lexicon, with small edit distance, to address the spelling variation problem is not feasible since 928 term pairs in the lexicon already have an edit distance of 1 (24% of the terms). Therefore, if we were to allow matching between terms and the dictionary even if they have an edit distance of 1 would introduce a large number of false positives or ambiguous matches.

To address this problem of low recall, we present a method for semi-automatic acquisition of transliterated term pairs using an English medical lexicon and our corpus. We will show that using this enhanced lexicon improves the performance on the term extraction task significantly.

Most previous work concerning transliterations focused on transliteration pair acquisition, *i.e.*, recognizing that two words (source, target) are equivalent, as one is a transliteration of the other. Transliteration pair acquisition includes two sub-tasks: recognizing that a lexeme contains transliteration and

finding the equivalent word in the source language (Knight and Graehl, 1998, Al-Onaizan and Knight, 2002). Another approach is using comparable corpora for detecting word pairs (Klementiev and Roth 2006). Such a parallel corpus is not available for medical Hebrew.

The first task, recognizing a transliterated word, is language dependent. It is fairly simple in languages such as Japanese in which transliterations are written in a different script than other Japanese words and are, therefore, easily identifiable. In other languages, such as Korean, Arabic and Hebrew, deciding which word needs to be back-transliterated is more complex. (Oh and Choi, 2000) suggested a method for Korean, based on supervised naïve Bayesian learning of phonemes and their combination in transliterated words and original Korean words. This method required manual tagging of the syllables in 1,900 documents as either Korean or foreign. (Baker and Brew, 2008) reported an accuracy of 96% in Korean, with a regression model trained on automatically generated data using phonetic rules instead of a manually tagged dataset.

To recognize transliterations in Arabic, (Nwesri et al., 2006) compared a lexicon-based approach with a supervised letter N-gram learning approach, suggested by (Cavnar and Trenkle, 1994), and a method based on recognizing Arabic specific patterns. The lexicon-based approach was most successful, augmented by heuristic rules, and resulted in precision of 47.7% and recall of 57.2%.

(Goldberg and Elhadad, 2008) developed a method for transliteration recognition in Hebrew based on an N-gram letter model. The method created a training set from a pronunciation dictionary automatically, thus the method is mostly unsupervised. Before applying the n-gram classifier, agglutinated affixes were manually removed from the words. This method achieved an F-Measure of 79% when assisted by a lexicon.

We extended this method and obtained significant performance improvement by combining morphological analysis and segmentation in the process of transliteration identification instead of manual segmentation as done by (Goldberg and Elhadad, 2008). The training set is created for domain specific words transliterated using pronunciation information from Miriam Webster Medical Dictionary.

For cross validation of this method, we used another domain specific corpus of gossip news from the Walla! website. Transliterations are common in both the medical and gossip domains: 8.5% of the word types in the medical domain are transliterations and 9% in the gossip domain. In the medical corpus, 4.5% of word instances are transliterations. Our method of transliteration recognition produces an F-measure of 93% for the medical domain and 94% for the gossip news domain.

We approach the second task, detecting the source language word (i.e. that "גאנציקלוויר" and "gunciclovir" are the same word), using the tokens identified as transliterations in the entire corpus. 10,000 suspect tokens were extracted. English medical lexicons were used to produce transliterations (without pronunciation data, up to 100K possible transliterations were created for each word). Suspect tokens were compared to the produced transliterations and a lexicon of transliterated pairs was created. False positives were manually removed from the lexicon, leaving 2,400 pairs (25% coverage of suspect tokens). 1,400 of these terms are not in the original lexicon adding ~20% more terms when combined. A useful property of the acquired dictionary is that it includes a link to the original English term in UMLS.

Evaluation on the task of term extraction (detecting terms from a medical lexicon in the documents) using the extended dictionary we have developed, improved recall from 33% to 39.6% and precision from 89% to 91.3% over the baseline.

References

Jain NL, Knirsch CA, Friedman C, Hripcsak G. (1996)Identification of suspected tuberculosis patients based on Natural language processing of chest radiograph reports. *American Medical Informatics Association* pp.542.

Simone Teufel and Noémie Elhadad. *Collection and Linguistic Processing of a Large-scale Corpus of Medical Articles*. 2002. LREC, pp. 1214-1218.
Adler, M. & Elhadad, M. (2006) An unsupervised morpheme-based hmm for hebrew morphological disambiguation. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* pp. 665-672).
Young-Suk, L., Kishore, P., Salim, R., Ossama, E. & Hany, H. (2003) Language model based arabic word segmentation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. Association for Computational Linguistics, Sapporo, Japan.
Knight, K. & Graehl, J. (1998). Machine transliteration. *Computational Linguistics* 24: 599-612.
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, pp. 400-408. Association for Computational Linguistics, Philadelphia, Pennsylvania.
Klementiev A, Roth D. (2006) Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. *Association for Computational Linguistics;* 2006: 817-824
Oh, J. & Choi, K. (2000) An English-Korean transliteration model using pronunciation and contextual rules.
Baker, K. & Brew, C. (2008) Statistical Identification of English Loanwords in Korean Using Automatically Generated Training Data. In: LREC 2008.
Nwesri, A., Tahaghoghi, S. & Scholer, F. (2006) Capturing out-of-vocabulary words in Arabic text. pp. 258-266).
Cavnar, W. & Trenkle, J. (1994). N-gram-based text categorization. *Ann Arbor MI* 48113: 4001.
Goldberg, Y. & Elhadad, M. (2008). Identification of transliterated foreign words in Hebrew script. *Lecture Notes in Computer Science* 4919: 466.
Dunning, T. (1994). Statistical identification of language. *Computing Research Laboratory Technical Memo MCCS*: 94-273.