Soundex-based Translation Correction in Urdu–English Cross-Language Information Retrieval

Manaal Faruqui
Computer Science & Engg.
Indian Institute of Technology
Kharagpur, India
manaalfar@gmail.com

Prasenjit Majumder
Computer Science & Engg.
DAIICT Gandhinagar
Gandhinagar, India
p_majumder@daiict.ac.in

Sebastian Padó
Computational Linguistics
Heidelberg University
Heidelberg, Germany
pado@cl.uni-heidelberg.de

Abstract

Cross-language information retrieval is difficult for languages with few processing tools or resources such as Urdu. An easy way of translating content words is provided by Google Translate, but due to lexicon limitations named entities (NEs) are transliterated letter by letter. The resulting NEs errors (*zynydyny zdn* for *Zinedine Zidane*) hurts retrieval. We propose to replace English non-words in the translation output. First, we determine phonetically similar English words with the Soundex algorithm. Then, we choose among them by a modified Levenshtein distance that models correct transliteration patterns. This strategy yields an improvement of 4% MAP (from 41.2 to 45.1, monolingual 51.4) on the FIRE-2010 dataset.

1 Introduction

Cross-language information retrieval (CLIR) research is the study of systems that accept queries in one language and return text documents in a different language. CLIR is of considerable practical importance in countries with many languages like India. One of the most widely used languages is Urdu, the official language of five Indian states as well as the national language of Pakistan. There are around 60 million speakers of Urdu – 48 million in India and 11 million in Pakistan (Lewis, 2009).

Despite this large number of speakers, NLP for Urdu is still at a fairly early stage (Hussain, 2008). Studies have been conducted on POS tagging (Sajjad and Schmid, 2009), corpus construction (Becker and Riaz, 2002), word segmentation (Durrani and Hussain, 2010), lexicographic sorting (Hussain et al., 2007), and information extraction (Mukund et al., 2010). Many other processing tasks are still missing, and the size of the Urdu internet is minuscule compared to English and other major languages, making Urdu a prime candidate for a CLIR source language.

A particular challenge which Urdu poses for CLIR is its writing system. Even though it is a Central Indo-Aryan language and closely related to Hindi, its development was shaped predominantly by Persian and Arabic, and it is written in Perso-Arabic script rather than Devanagari. CLIR with a target language that uses another script needs to transliterate (Knight and Graehl, 1998) any material that cannot be translated (typically out-of-vocabulary items like Named Entities). The difficulties of Perso-Arabic in this respect are (a), some vowels are represented by letters which are also consonants and (b), short vowels are customarily omitted. For example, in وِنونا (*Winona*) the first و is used for the *W* but the second is used for *O*. Also the *i* sound is missing after و (*W*).

In this paper, we consider Urdu–English CLIR. Starting from a readily available baseline (using Google Translate to obtain English queries), we show that transliteration of Named Entities, more specifically missing vowels, is indeed a major factor in wrongly answered queries. We reconstruct missing vowels in an unsupervised manner through an approximate string matching procedure based on phonetic similarity and orthographic similarity by using Soundex code (Knuth, 1975) and Levenshtein distance (Gusfield, 1997) respectively, and find a clear improvement over the baseline.

2 Translation Strategies for Urdu–English

We present a series of strategies for translating Urdu queries into English so that they can be pre-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 25–29,
Chiang Mai, Thailand, November 8-12, 2011.

sented to a monolingual English IR system that works on some English document collection. Inspection of the strategies' errors led us to develop a hierarchy of increasingly sophisticated strategies.

2.1 Baseline model (GTR)

As our baseline, we aimed for a model that is state-of-the-art, freely available, and can be used by users without the need for heavy computational machinery. We decided to render the Urdu query into English with the Google Translate web service.[1]

2.2 Approximate Matching (GTR+SoEx)

Google Translate appears to have a limited Urdu lexicon. Words that are out of vocabulary (OOV) are transliterated letter by letter into the Latin alphabet. Without an attempt to restore short (unwritten) vowels, these match the actual English terms only very rarely. For example, *Singur*, the name of a village in India gets translated to *Sngur*.

To address this problem, we attempt to map these incomplete transliterations onto well-formed English words using approximate string matching. We use Soundex (Knuth, 1975), an algorithm which is normally used for "phonetic normalization". Soundex maps English words onto their first letter plus three digits which represent equivalence classes over consonants, throwing away all vowels in the process. For example, *Ashcraft* is mapped onto A261, where 2 stands for the "gutturals" and "sibilants" S and K, 6 for R, and 1 for the "labiodental" F. All codes beyond the first three are ignored. The same soundex code would be assigned, for example, to *Ashcroft*, *Ashcrop*, or even *Azaroff*. The two components which make Soundex a well-suited choice for our purposes are exactly (a), the forming of equivalence classes over consonants, which counteracts variance introduced by one-to-many correspondences between Latin and Arabic letters; and (b), the omission of vowels.

Specifically, we use Soundex as a hash function, mapping all English words from our English document collection onto their Soundex codes. The GTR+SoEx model then attempts to correct all words in the Google Translate output by replacing them with the English word sharing the same Soundex code that has the highest frequency in the English document collection.

[1] http://translate.google.com. All queries were translated in the first week of January 2011.

2.3 NER-centered Approximate Matching (GTR+SoExNER)

An analysis of the output of the GTR+SoEx model showed that the model indeed ensured that all words in the translation were English words, but that it "overcorrected", replacing correctly translated, but infrequent, English words by more frequent words with the same Soundex code. Unfortunately, Google Translate does not indicate which words in its output are out-of-vocabulary.

Recall that our original motivation was to improve coverage specifically for out-of-vocabulary words, virtually all of which are Named Entities. Thus, we decided to apply Soundex matching only to NEs. As a practical and simple way of identifying malformed NEs, we considered those words in the Google Translate output which did not occur in the English document base at all (i.e., which were "non-words"). We manually verified that this heuristic indeed identified malformed Named Entities in our experimental materials (see Section 3 below for details). We found a recall of 100% (all true NEs were identified) and a precision of 96% (a small number of non-NEs was classified as NEs).

The GTR+SoExNER strategy applies Soundex matching to all NEs, but not to other words in the Google Translate output.

2.4 Disambiguation (GTR+SoExNER+LD (mod))

Generally, a word that has been wrongly transliterated from Urdu maps onto the same Soundex code as several English words. The median number of English words per transliteration is 7. This can be seen as a sort of ambiguity, and the strategy adopted by the previous models is to just choose the most frequent candidate, similar to the "predominant" sense baseline in word sense disambiguation (McCarthy et al., 2004). We found however that the most frequent candidate is often wrong, since Soundex conflates fairly different words (cf. Section 2.2). For example, *Subhas*, the first name of an Indian freedom fighter, receives the soundex code *S120* but it is mapped onto the English term *Space* ($freq$=7243) instead of *Subhas* ($freq$=2853).

We therefore experimented with a more informed strategy that chooses the English candidate based on two variants of Levenshtein distance. The first model, GTR+SoExNER+LD, uses standard Levenshtein distance with a cost of 1 for

each insertion, deletion and substitution. Our final model, GTR+SoExNER+LDmod uses a modified version of Levenshtein distance which is optimized to model the correspondences that we expect. Specifically, the addition of vowels and the replacement of consonants by vowels come with no cost, to favour the recovery of English vowels that are unexpressed in Urdu or expressed as consonants (cf. Section 1). Thus, the LDmod between *zdn* and *zidane* would be Zero.

3 Experimental Setup

Document Collection and Queries Our experiments are based on the FIRE-2010[2] English data, consisting of documents and queries, as our experimental materials. The document collection consists of about 124,000 documents from the English-language newspaper "The Telegraph India"[3] from 2004-07. The average length of a document was 40 words. The FIRE query collection consists of 50 English queries which were of the same domain as that of the document collection. The average number of relevant documents for a query was 76 (with a minimum of 13 and a maximum of 228).

The first author, who has an advanced knowledge of Urdu, translated the English FIRE queries manually into Urdu. One of the resulting Urdu query is shown in Table 1, together with the Google translations back into English (GTR) which form the basis of the CLIR queries in the simplest model. Every query has a *title*, and a *description*, both of which we used for retrieval. The bottom row (*entity*) shows the Translate output and from the best model (Soundex matching with modified Levenshtein distance). The bold-faced terms correspond to names that are corrected successfully, increasing the query's precision from 49% to 86%.

Cross-lingual IR setup We implemented the models described in Section 2, using the Terrier IR engine (Ounis et al., 2006) for retrieval from the FIRE-2010 English document collection. We used the PL2 weighting model with the term frequecy normalisation parameter of 10.99. The document collection and the queries were stemmed using the Porter Stemmer (Porter, 1980). We applied all translation strategies defined in Section 2 as *query expansion* modules that enrich the Google Translate output with new relevant query terms. In a pre-experiment, we experimented with adding either only the single most similar term for each OOV item (1-best) or the best n terms (n-best). We consistently found better results for 1-best and report results for this condition only.

Monolingual model We also computed a monolingual English model which did not use the translated Urdu queries but the original English ones instead. The result for this model can be seen as an upper bound for Urdu-English CLIR models.

Evaluation We report two evaluation measures. The first one is Mean Average Precision (MAP), an evaluation measure that is highest when all correct items are ranked at the top (Manning et al., 2008). MAP measures the global quality of the ranked document list; however improvements in MAP could result from an improved treatment of marginally relevant documents, while it is the quality of the top-ranked documents that is most important in practice and correlates best with extrinsic measures (Scholer and Turpin, 2009). Therefore we also consider P@5, the precision of the five top-ranked documents.

4 Results and Discussion

Table 2 shows the results of our experiments. Monolingual English retrieval achieves a MAP of 51.4, while the CLIR baseline (Google Translate only – GTR) is 41.3. We expect the results of our experiments to fall between these two extremes.

We first extend the baseline model with Soundex matching for all terms in the title and description (GTR+SoEx), we actually obtain a result way below the baseline (MAP=36.7). The reason is that, as discussed in Section 2.2, Soundex is too coarse-grained for non-NEs, grouping words such as *red* and *road* into the same equivalence class, thus pulling in irrelevant terms. This analysis is supported by the observation, mentioned above, that 1-best always performs better than n-best.

We are however able to obtain a clear improvement of about 1.5% absolute by limiting Soundex matching to automatically identified Named Entities, up to MAP=43.0 (GTR+SoExNER). However, this model still relies completely on frequency for choosing among competitors with the same Soundex code, leading to errors like the *Subhas*/*Space* mixup discusssed in Section 2.4. The use of Levenshtein distance, representing a more informed manner of disambiguation, makes

[2] http://www.isical.ac.in/~fire/2010/data_download.html

[3] http://www.telegraphindia.com/

title UR	زینیدین زیذان کا ورلڈ کپ میں سر سے مارنے کا واقعہ
title EN (GTR)	Zynydyny zydan World Cup head butt incident
desc UR	ایسے دستاویزات کو تلاش کریں جس میں زیذان نے ماتیرّی کو سر سے ورلڈ کپ ۲۰۰۶ کے فاءینل میں مارا جب اتالوی نے زیذان کے خلاف ناگوار باتیں بولیں
desc EN (GTR)	Find these documents from public opinion zdn to mtrzzy, from Italian to zydan about offensive comments, World Cup finals in 2006 head to kill incidents are mentioned
entity EN (GTR)	Zynydyny Zydan zdn Mtrzzy
entity (GTR+SoExNER+LDmod)	**zinedine** zaydan **zidane materazzi**

Table 1: A sample query

Model	MAP	P@5
GTR	41.3	62.4
GTR+SoEx	36.7	59.2
GTR+SoExNER	43.0	62.4
GTR+SoExNER+LD	45.0	65.2
GTR+SoExNER+LDmod	**45.3**	**65.6**
Monolingual English	51.4	71.6

Table 2: Results for Urdu-English CLIR models on the FIRE 2010 collection (Mean Average Precision and Precision of top five documents)

a considerable difference, and leads to a final MAP of 45.33 or about 4% absolute increase for the (`GTR+SoExNER+LDmod`) model. A bootstrap resampling analysis (Efron and Tibshirani, 1994) confirmed that the difference between `GTR+SoExNER+LDmod` and `GTR` model is significant ($p<0.05$). All models are still significantly worse than the monolingual English model.

The P@5 results are in tandem with the MAP results for all models, showing that the improvement which we obtain for the best model leads to top-5 lists whose precision is on average more than 3% better than the baseline top-5 lists. This difference is not significant, but we attribute the absence of significance to the small sample size (50 queries).

In a qualitative analysis, we found that many remaining low-MAP queries still suffer from missing or incorrect Named Entities. For example, *Noida* (an industrial area near New Delhi), was transliterated to *Nuydh* and then incorrectly modified to *Nidhi* (an Indian name). This case demonstrates the limits of our method which cannot distinguish well among NEs which differ mainly in their vowels.

5 Related Work

There are several areas of related work. The first is IR in Urdu, where monolingual work has been done (Riaz, 2008). However, to our knowledge, our study is the first one to address Urdu CLIR. The second is machine transliteration, which is a widely researched area (Knight and Graehl, 1998) but which usually requires some sort of bilingual resource. Knight and Graehl (1998) use 8000 English-Japanese place name pairs, and Mandal et al. (2007) hand-code rules for Hindi and Bengali to English. In contrast, our method does not require any bilingual resources. Finally, Soundex codes have been applied to Thai-English CLIR (Suwanvisat and Prasitjutrakul, 1998) and Arabic name search (Aqeel et al., 2006). They have also been found useful for indexing Named Entities (Raghavan and Allan, 2004; Kondrak, 2004) as well as IR more generally (Holmes and McCabe, 2002).

6 Conclusion

In this paper, we have considered CLIR from Urdu into English. With Google Translate as translation system, the biggest hurdle is that most named entities are out-of-vocabulary items and transliterated incorrectly. A simple, completely unsupervised postprocessing strategy that replaces English non-words by phonetically similar words with minimal edit distance is able to recover almost half of the loss in MAP that the cross-lingal setup incurs over a monolingual English one. Directions for future work include monolingual query expansion in Urdu to improve the non-NE part of the query and training a full Urdu-English transliteration system.

Acknowledgments We thank A. Tripathi and H. Sajjad for invaluable discussions and suggestions.

References

Syed Uzair Aqeel, Steve Beitzel, Eric Jensen, David Grossman, and Ophir Frieder. 2006. On the development of name search techniques for Arabic. *Journal of the American Society for Information Science and Technology*, 57(6):728–739.

Dara Becker and Kashif Riaz. 2002. A study in urdu corpus construction. In *Proceedings of the 3rd COLING workshop on Asian language resources and international standardization*, pages 1–5, Taipei, Taiwan.

Nadir Durrani and Sarmad Hussain. 2010. Urdu word segmentation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 528–536, Los Angeles, California.

Bradley Efron and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall.

Dan Gusfield. 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, Cambridge, UK.

David Holmes and M. Catherine McCabe. 2002. Improving precision and recall for soundex retrieval. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, pages 22–27, Washington, DC, USA.

Sarmad Hussain, Sana Gul, and Afifah Waseem. 2007. Developing lexicographic sorting: An example for urdu. *ACM Transactions on Asian Language Information Processing*, 6(3):10:1–10:17.

Sarmad Hussain. 2008. Resources for urdu language processing. In *Proceedings of the Workshop on Asian Language Resources at IJCNLP 2008*, Hyderabad, India.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Donald E. Knuth. 1975. *Fundamental Algorithms*, volume III of *The Art of Computer Programming*. Addison-Wesley, Reading, MA.

Grzegorz Kondrak. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *In Proceedings of the International Conference on Computational Linguistics*, pages 952–958, Geneva, Switzerland.

M. Paul Lewis, editor. 2009. *Ethnologue – Languages of the World*. SIL International, 16th edition.

Debasis Mandal, Mayank Gupta, Sandipan Dandapat, Pratyush Banerjee, and Sudeshna Sarkar. 2007. Bengali and Hindi to English CLIR evaluation. In *Proceedings of CLEF*, pages 95–102.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 1st edition.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

Smruthi Mukund, Rohini Srihari, and Erik Peterson. 2010. An information-extraction system for urdu—a resource-poor language. *ACM Transactions on Asian Language Information Processing*, 9:15:1–15:43.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval*, Seattle, WA, USA.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Hema Raghavan and James Allan. 2004. Using soundex codes for indexing names in ASR documents. In *Proceedings of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, pages 22–27, Boston, MA.

Kashif Riaz. 2008. Baseline for Urdu IR evaluation. In *Proceeding of the 2nd ACM workshop on Improving non-English web searching*, pages 97–100, Napa, CA, USA.

Hassan Sajjad and Helmut Schmid. 2009. Tagging Urdu text with parts of speech: A tagger comparison. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 692–700, Athens, Greece.

Falk Scholer and Andrew Turpin. 2009. Metric and relevance mismatch in retrieval evaluation. In *Information Retrieval Technology*, volume 5839 of *Lecture Notes in Computer Science*, pages 50–62. Springer.

Prayut Suwanvisat and Somboon Prasitjutrakul. 1998. Thai-English cross-language transliterated word retrieval using soundex technique. In *Proceesings of the National Computer Science and Engineering Conference*, Bangkok, Thailand.