Analyzing Urdu Social Media for Sentiments using Transfer Learning with Controlled Translations

Smruthi Mukund
CEDAR, Davis Hall, Suite 113
University at Buffalo, SUNY, Buffalo, NY
smukund@buffalo.edu

Rohini K Srihari
CEDAR, Davis Hall, Suite 113
University at Buffalo, SUNY, Buffalo, NY
rohini@cedar.buffalo.edu

Abstract

The main aim of this work is to perform sentiment analysis on Urdu blog data. We use the method of structural correspondence learning (SCL) to transfer sentiment analysis learning from Urdu newswire data to Urdu blog data. The pivots needed to transfer learning from newswire domain to blog domain is not trivial as Urdu blog data, unlike newswire data is written in Latin script and exhibits code-mixing and code-switching behavior. We consider two oracles to generate the pivots. 1. Transliteration oracle, to accommodate script variation and spelling variation and 2. Translation oracle, to accommodate code-switching and code-mixing behavior. In order to identify strong candidates for translation, we propose a novel part-of-speech tagging method that helps select words based on POS categories that strongly reflect code-mixing behavior. We validate our approach against a supervised learning method and show that the performance of our proposed approach is comparable.

1 Introduction

The ability to break language barriers and understand people's feelings and emotions towards societal issues can assist in bridging the gulf that exists today. Often emotions are captured in blogs or discussion forums where writers are common people empathizing with the situations they describe. As an example, the incident where a cricket team visiting Pakistan was attacked caused widespread anguish among the youth in that country who thought that they will no longer be able to host international tournaments. The angry emotion was towards the failure of the government to provide adequate protection for citizens and visitors. Discussion forums and blogs on cricket, mainly written by Pakistani cricket fans, around the time, verbalized this emotion. Clearly analyzing blog data helps to estimate emotion responses to domestic situations that are common to many societies.

Traditional approaches to sentiment analysis require access to annotated data. But facilitating such data is laborious, time consuming and most importantly fail to scale to new domains and capture peculiarities that blog data exhibits; 1. spelling variations and 2. code mixing and code switching. 3. script difference (Nastaliq vs Latin script). In this work, we present a new approach to polarity classification of code-mixed data that builds on a theory called structural correspondence learning (SCL) for domain adaptation. This approach uses labeled polarity data from the base language (in this case, Urdu newswire data - source) along with two simple oracles that provide one-one mapping between the source and the target data set (Urdu blog data).

Subsequent sections are organized as follows. Section 2 describes the issues seen in Urdu blog data followed by section 3 that explains the concept of structural correspondence learning. Section 4 details the code mixing and code switching behavior seen in blog data. Section 5 describes the statistical part of speech (POS) tagger developed for blog data required to identify mixing patterns followed by the sentiment analysis model in section 6. We conclude with section 7 and briefly outline analysis and future work in section 8.

Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 1–8,
Montréal, Canada, June 7, 2012.

2 Urdu Blog Data

Though non-topical text analysis like emotion detection and sentiment analysis, have been explored mostly in the English language, they have also gained some exposure in non-English languages like Urdu (Mukund and Srihari, 2010), Arabic (Mageed *et al.,* 2011) and Hindi (Joshi and Bhattacharya, 2012). Urdu newswire data is written using Nastaliq script and follows a relatively strict grammatical guideline. Many of the techniques proposed either depend heavily on NLP features or annotated data. But, data in blogs and discussion forums especially written in a language like Urdu cannot be analyzed by using modules developed for Nastaliq script for the following reasons; (1) the tone of the text in blogs and discussion forums is informal and hence differs in the grammatical structure (2) the text is written using Latin script (3) the text exhibits code mixing and code switching behavior (with English) (4) there exists spelling errors which occur mostly due to the lack of predefined standards to represent Urdu data in Latin script.

Urdish (Urdu blog data) is the term used for Urdu, which is (1) written either in Nastaliq or Latin script, and (2) contains several English words/phrases/sentences. In other words, Urdish is a name given to a language that has Urdu as the base language and English as the seasoning language. With the wide spread use of English keyboards these days, using Latin script to encode Urdu is very common. Data in Urdish is never in pure Urdu. English words and phrases are commonly used in the flow integrating tightly with the base language. Table 1 shows examples of different flavors in which Urdu appears in the internet.

Different Forms of Data	**Main Issues**	**Example Sentence**
1. Urdu written in Nastaliq	1. Lack of tools for basic operations such as segmentation and diacritic restoration 2. Lack of sufficient annotated data for POS and NE tagging 3. Lack of annotated data for more advanced NLP	فوجی جوانوں کو کئی لوگوں سے غصہ آگیا *[The soldiers were angry with a lot of people]*
2. Urdu written in ASCII (English)	1. Several variations in spellings that need to be normalized 2. No normalization standards 3. Preprocessing modules needed if tools for Urdu in Nastaliq are to be used 4. Developing a completely new NLP framework needs annotated data	*Wo Mulk Jisko Hum nay 1000000 logoon sey zayada Loogoon ki Qurbanian dey ker hasil kia usi mulk main yai kaisa waqt a gay hai ?* *[Look at what kind of time the land that had 1000000's of people sacrifice their lives is experiencing now]*
3. Urdish written in Nastaliq	1. No combined parser that deals with English and Urdu simultaneously 2. English is written in Urdu but with missing diacritics	ٹی وی سٹیشن میں فون پر فون آنے لگے *[the phones rang one after the other in the TV station]*
4. Urdish written in ASCII(English)	**1. No combined parser that deals with English and Urdu simultaneously** **2. Issue of spelling variations that need to be normalized**	**Afsoos key baat hai . kal tak jo batain Non Muslim bhi kartay hoay dartay thay abhi this man has brought it out in the open.** ***[It is sad to see that those words that even a non muslim would fear to utter till yesterday, this man had brought it out in the open]***

Table 1: Different forms of using Urdu language on the internet

Blog data follows the order shown in example 4 of table 1. Such a code-switching phenomenon is very common in multilingual societies that have significant exposure to English. Other languages exhibiting similar behaviors are Hinglish (Hindi and English), Arabic with English and Spanglish (Spanish with English).

3 Structural Correspondence Learning

For a problem where domain and data changes requires new training and learning, resorting to classical approaches that need annotated data becomes expensive. The need for domain adaptation arises in many NLP tasks – part of speech tagging, semantic role labeling, dependency parsing, and sentiment analysis and has gained high visibility in the recent years (Daume III and Marcu, 2006; Daume III *et al.*, 2007; Blitzer *et al.*, 2006, Prettenhofer and Stein *et al.*, 2010). There exists two main approaches; supervised and semi-supervised.

In the supervised domain adaptation approach along with labeled source data, there is also access to a small amount of labeled target data. Techniques proposed by Gildea (2001), Roark and Bacchiani (2003), Daume III (2007) are based on the supervised approach. Studies have shown that baseline approaches (based on source only, target only or union of data) for supervised domain adaption work reasonably well and beating this is surprisingly difficult (Daume III, 2007).

In contract, the semi supervised domain adaptation approach has access to labeled data only in the source domain (Blitzer *et al.*, 2006; Dredze *et al.*, 2007; Prettenhofer and Stein *et al.*, 2010). Since there is no access to labeled target data, achieving baseline performance exhibited in the supervised approach requires innovative thinking.

The method of structural correspondence learning (SCL) is related to the structural learning paradigm introduced by Ando and Zhang (2005). The basic idea of structural learning is to constrain the hypothesis space of a learning task by considering multiple different but related tasks on the same input space. SCL was first proposed by Blitzer *et al.*, (2006) for the semi supervised domain adaptation problem and works as follows (Shimizu and Nakagawa, 2007).

1. A set of pivot features are defined on unlabeled data from both the source domain and the target domain
2. These pivot features are used to learn a mapping from the original feature spaces of both domains to a shared, low-dimensional real–valued feature space. A high inner product in this new space indicates a high degree of correspondence along that feature dimension
3. Both the transformed and the original features in the source domain are used to train a learning model
4. The effectiveness of the classifier in the source domain transfers to the target domain based on the mapping learnt

This approach of SCL was applied in the field of cross language sentiment classification scenario by Prettenhofer and Stein (2010) where English was used as the source language and German, French and Japanese as target languages. Their approach induces correspondence among the words from both languages by means of a small number of pivot pairs that are words that process similar semantics in both the source and the target languages. The correlation between the pivots is modeled by a linear classifier and used as a language independent predictor for the two equivalent classes. This approach solves the classification problem directly, instead of resorting to a more general and potentially much harder problem such as machine translation.

The problem of sentiment classification in blog data can be considered as falling in the realm of domain adaptation. In this work, we approach this problem using SCL tailored to accommodate the challenges that code-mixed data exhibits. Similar to the work done by Prettenhofer and Stein (2010), we look at generating pivot pairs that capture code-mixing and code-switching behavior and language change.

4 Code Switching and Code Mixing

Code switching refers to the switch that exists from one language to another and typically involves the use of longer phrases or clauses of another language while conversing in a totally different base language. Code mixing, on the other hand, is a phenomenon of mixing words and other smaller units of one language into the structure of another language. This is mostly inter-sentential.

In a society that is bilingual such as that in Pakistan and India, the use of English in the native language suggests power, social prestige and the status. The younger crowd that is technologically well equipped tends to use the switching phenomenon in their language, be it spoken or written. Several blogs, discussion forums, chat rooms *etc*. hold information that is expressed is intensely code mixed. Urdu blog data exhibits mix of Urdu language with English.

There are several challenges associated with developing NLP systems for code-switched languages. Work done by Kumar (1986) and Sinha & Thakur, (2005) address issues and challenges associated with Hinglish (Hindi – English) data. Dussias (2003) and Celia (1997) give an overview of the behavior of code switching occurring in Spanish - Spanglish. This phenomenon can be seen in other languages like Kannada and English, German and English. Rasul (2006) analyzes the linguistic patterns occurring in Urdish (Urdu and English) language. He tries to quantize the extent to which code-mixing occurs in media data, in particular television. Most of his rules are based on

what is proposed by Kachru (1978) for Hinglish and has a pure linguistic approach with manual intervention for both qualitative and quantitative analysis.

Several automated techniques proposed for Hinglish and Spanglish are in the context of machine translation and may not be relevant for a task like information retrieval since converting the data to one standardized form is not required. A more recent work was by Goyal *et al.*, (2003) where they developed a bilingual parser for Hindi and English by treating the code mixed language as a completely different variety. However, the credibility of the system depends on the availability of WordNet[1].

4.1 Understanding Mixing Patterns

Performing analysis on data that exhibit code-switching has been attempted by many across various languages. Since the Urdu language is very similar to Hindi, in this section we discuss the code-mixing behavior based on a whole battery of work done by researchers in the Hindi language.

Researchers have studied the behavior of the mixed patterns and generated rules and constraints on code-mixing. The study of code mixing with Hindi as the base language is attempted by Sinha and Thakur (2005) in the context of machine translation. They categorize the phenomenon into two types based on the extent to which mixing happens in text in the context of the main verb. Linguists such as Kachru (1996) and Poplack (1980) have tried to formalize the terminologies used in this kind of behavior. Kumar (1986) says that the motivation for assuming that the switching occurs based on certain set of rules and constraints are based on the fact that users who use this can effectively communicate with each other despite the mixed language. In his paper he proposes a set of rules and constraints for Hindi-English code switching. However, these rules and constraints have been countered by examples proposed in the literature (Agnihotri, 1998). This does not mean that researchers earlier had not considered all the possibilities. It only means that like any other language, the language of code-mixing is evolving over time but at a very fast pace.

One way to address this problem of code-mixing and code switching for our task of sentiment analysis in blog data is rely on predefined rules to identify mixed words. But this can get laborious and the rules may be insufficient to capture the latest behavior. Our approach is to use a statistical POS model to determine part of speech categories of words that typically undergo such switches.

5 Statistical Part of Speech Tagger

Example 5.1 showcases a typical sentence seen in blog data. Example 5.2 shows the issue with spelling variations sometimes that occur in the same sentence

Example 5.1: *Otherwise humara bhi wohi haal hoga jo is time Palestine, Iraq, Afghanistan wagera ka hai ~ Otherwise our state will also be like what is in Palestine, Iraq, Afghanistan etc. are experiencing at this time*

Example 5.2: *Shariyat **ke** aitebaar se bhi ghaur kia jaey tu aap ko ilm ho jaega **key** **joh** haraam khata **hai** uska dil kis tarhan ka hota **hey** ~ If you look at it from morals point of you too you will understand the heart of people who cheat*

A statistical POS tagger for blog data has to take into consideration spelling variations, mixing patterns and script change. The goal here is not to generate a perfect POS tagger for blog data (though the idea explained here can be extended for further improvisation) but to be able to identify POS categories that are candidates for switch and mix. The basic idea of our approach is as follows

1. Train Latin script POS tagger (LS tagger) on pure Urdu Latin script data (Example 2 in table 1 – using Urdu POS tag set, Muaz *et al.*, 2009)
2. Train English POS tagger on English data (based on English tag sets, Santorini, 1990)
3. Apply LS tagger and English tagger on Urdish data and note the confidence measures of the applied tags on each word
4. Use confidence measures, LS tags, phoneme codes (to accommodate spelling variations) as features to train a new learning model on Urdish data
5. Those words that get tagged with the English tagset are potential place holders for mixing patterns

Word	**Act**	**Eng**	**LS Urdu**	**Urd CM**	**Eng CM**
and	CC	**CC**	NN	0.29	0.99
most	RB	**RB**	VM	0.16	0.83
important	JJ	**JJ**	VAUX	0.08	0.97
thing	NN	**NN**	CC	0.06	0.91

[1] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

Zardari	NNP	NNP	**NN**	0.69	0.18
ko	PSP	NNP	**PSP**	0.99	0.28
shoot	VB	NNP	**JJ**	0.54	0.29
ker	NN	NNP	**NN**	0.73	0.29
dena	VM	NNP	**VM**	0.83	0.29
chahiya	VAUX	NNP	**VAUX**	0.98	0.21
.	SYM	.	**SYM**	0.99	0.99

Table 2. POS tagger with confidence measures

The training data needed to develop LS tagger for Urdu is obtained from Hindi. IIIT POS annotated corpus for Hindi contains data in the SSF format (Shakti Standard Format) (Bharati, 2006). This format tries to capture the pronunciation information by assigning unique English characters to Hindi characters. Since this data is already in Latin script with each character capturing a unique pronunciation, changing this data to a form that replicates chat data using heuristic rules is trivial. However, this data is highly sanskritized and hence need to be changed by replacing Sanskrit words with equivalent Urdu words. This replacement is done by using online English to Urdu dictionaries (www.urduword.com and www.hamariweb.com). We have succeeded in replacing 20,000 pure Sanskrit words to Urdu by performing a manual lookup. The advantage with this method is that

1. The whole process of setting up annotation guidelines and standards is eliminated.
2. The replacement of pure Hindi words with Urdu words in most cases is one-one and the POS assignment is retained without disturbing the entire structure of the sentence.

Our training data now consists of Urdu words written in Latin script. We also generate phonemes for each word by running the phonetic model. A POS model is trained using CRF (Lafferty, 2001) learning method with current word, previous word and the phonemes as features. This model called the Latin Script (LS) POS model has an F-score of 83%.

English POS tagger is the Stanford tagger that has a tagging accuracy of about 98.7%[2] .

5.1 Approach

Urdish blog data consists of Urdu code-mixed with English. Running simple Latin script based Urdu POS tagger results in 81.2% accuracy when POS tags on the entire corpus is considered and 52.3% accuracy on only the English words. Running English tagger on the entire corpus improves the POS tagging accuracy of English words to 79.2% accuracy. However, the tagging accuracy on the entire corpus reduces considerably – 55.4%. This indicates that identifying the language of the words will definitely improve tagging.

Identifying the language of the words can be done simply by a lexicon lookup. Since English words are easily accessible and more enriched, English Wordnet[3] makes a good source to perform this lookup. Running Latin script POS tagger and English tagger on the language specific words resulted in 79.82% accuracy for the entire corpus and 59.2% accuracy for English words. Clearly there is no significant gain in the performance. This is on account of English equivalent Urdu representation of words (e.g. *key* ~ their, *more* ~ peacock, *bat* ~ speak).

Since identifying the language explicitly yields less benefit, we showcase a new approach that is based on the confidence measures of the taggers. We first run the English POS tagger on the entire corpus. This tagger is trained using a CRF model. Scores that indicate the confidence with which this tagger has applied tags to each word in the corpus is also estimated (table 2). Next, the Latin script tagger is applied on the entire corpus and the confidence scores for the selected tags are estimated. So, for each word, there exist two tags, one from the English tagger and the other from the Latin script Urdish tagger along with their confidence scores. This becomes our training corpus.

The CRF learning model trained on the above corpus using features shown in table 3 generates a cross validation accuracy is 90.34%. The accuracy on the test set is 88.2%, clearly indicating the advantages of the statistical approach.

Features used to train Urdish POS tagger
Urdish word
POS tag generated by LS tagger
POS tag generated by English tagger
Confidence measure by LS tagger
Confidence measure by English tagger
Double metaphone value
Previous and next tags for English and Urdu
Previous and next words
Confidence priorities

Table 3. Features used to train the final POS tagger for Urdish data

[2] http://nlp.stanford.edu/software/tagger.shtml

[3] http://wordnet.princeton.edu/

Table 4 illustrates the POS categories used as potential pattern switching place holders

POS Category	Example
noun within a noun phrase	uski **life** par itna control acha nahi hai ~ its not good to control his life this much
Interjection	Comon Reema **yaar**! ~ Hey Man Reema! lol! ~ lol
Adjective	Yeh story bahut hi **scary** or **ugly** tha ~ This story was really scary and ugly
Adverb	Babra Shareef ki koi bhi film lagti hai, hum **definitely** dekhtai ~ I would definitely watch any movie of Babra Shareef
Gerund (tagged as a verb by English POS tagger)	Yaha **shooting** mana hai ~ shooting is prohibited here
Verb	Iss movie main I **dozed** ~ I slept through the movie
Verb	Afridi.. **Cool** off!

Table 4. POS categories that exhibit pattern switch

6 Sentiment Polarity Detection

The main goal of this work is to perform sentiment analysis in Urdu blog data. However, this task is not trivial owing to all the peculiarities that blog data exhibits. The work done on Urdu sentiment analysis (Mukund and Srihari, 2010) provided annotated data for sentiments in newswire domain. . Newspaper data make a good corpus to analyze different kinds of emotions and emotional traits of the people. They reflect the collective sentiments and emotions of the people and in turn the society to which they cater. When specific frames are considered (such as semantic verb frames) in the context of the triggering entities – *opinion holders* (entities who express these emotions) and *opinion targets* (entities towards whom the emotion is directed) - performing sentiment analysis becomes more meaningful and newspapers make an excellent source to analyze such phenomena (*Mukund et al.,* 2011). We use SCL to transfer sentiment analysis learning from this newswire data to blog data. Inspired by the work done by (Prettenhofer and Stein, 2010), we rely on oracles to generate pivot pairs. A pivot pair $\{w_S, w_T\}$ where $w_S \in V_S$ (the source language – Urdu newswire data) and $w_T \in V_T$ (the target language – Urdish data) should satisfy two conditions 1. high support and 2. high confidence, making sure that the pairs are predictive of the task.

Prettenhofer and Stein (2010) used a simple translation oracle in their experiments. However there exist several challenges with Urdish data that inhibits the use of a simple translation oracle.

1. Script difference in the source and target languages. Source corpus (Urdu) is written in Nastaleeq and the target corpus (Urdish) is written in ASCII
2. Spelling variations in roman Urdu
3. Frequent use of English words to express strong emotions

We use two oracles to generate pivot pairs.

The first oracle accommodates the issue with spelling variations. Each Urdu word is converted to roman Urdu using IPA (1999) guidelines. Using the double metaphone algorithm[4] phoneme code for the Urdu word is determined. This is also applied to Urdish data at the target end. Words that have the same metaphone code across the source and target languages are considered pivot pairs.

The second oracle is a simple translation oracle between Urdu and English. Our first experiment (experiment 1) is using words that belong to the adjective part of speech category as candidates for pivots. We augment this set to include words that belong to other POS categories shown in table 4 that exhibit pattern mixing (experiment 2).

6.1 Implementation

The feature used to train the learning algorithm is limited to unigrams. For linear classification, we use libSVM (Chang and Lin, 2011). The computational bottleneck of this method is in the SVD decomposition of the dense parameter matrix W. We set the negative values of W to zero to get a sparse representation of the matrix. For SVD computation the Lanczos algorithm provided by SVDLIBC[5] is employed. Each feature matrix used in libSVM is scaled between -1 and 1 and the final matrix for SVD is standardized to zero mean and unit variance estimated on $D_S \cup D_u$ (source subset and target subset).

6.2 Results

The domain of the source data set is limited to cricket and movies in order to ensure domain over-

[4] http://en.wikipedia.org/wiki/Double_Metaphone
[5] http://tedlab.mit.edu/~dr/SVDLIBC

lap between newswire data that we have and blog data. In order to benchmark the proposed technique, our baseline technique is based on the conventional method of supervised learning approach on annotated data. Urdish data set used for polarity classification contains 705 sentences written in ASCII format (example 6.1). This corpus is manually annotated by one annotator (purely based on intuition and does not follow any predefined annotation guidelines) to get 440 negative sentences and 265 positive sentences. The annotated corpus is purely used for testing and in this work considered as unlabeled data. A suitable linear kernel based support vector machine is modeled on the annotated data and a five-fold cross validation on this set gives an F-Measure of 64.3%.

Example 6.1:

General zia-ul-haq ke zamane mai qabayli elaqe Russia ke khilaf jang ka merkaz thea aur general Pervez Musharraf ke zamane mai ye qabayli elaqe Pakistan ke khilaf jang ka markaz ban gye . ~ negative

Our first experiment is based on using the second oracle for translations on only adjectives (most obvious choice for emotion words). We use 438 pivot pairs. The average F-measure for the performance is at 55.78% which is still much below the baseline performance of 64.3% if we had access to annotated data. However, the results show the ability of this method.

Our second experiment expands the power of the second oracle to provide translations to other POS categories that exhibit pattern switching. This increased the number of pivot pairs to 640. Increase in pivots improved the precision. Also we see significant improvement in the recall. The newly added pivots brought more sentences under the radar of the transfer model. The average F-Measure increased to 59.71%.

The approach can be further enhanced by improving the oracle used to select pivot features. One way is add more pivot pairs based on the correlation in the topic space across language domains (future work).

7 Conclusion

In this work we show a way to perform sentiment analysis in blog data by using the method of structural correspondence learning. This method accommodates the various issues with blog data such as spelling variations, script difference, pattern switching.

Precision (P %)	Recall (R %)	F-Measure (F %)
Phonemes (Roman Urdu)		
37.97	58.82	46.15
Metaphones based synonym mapping (adjectives)		
50.9	51	50.89
56.6	56.4	55.62
58.9	60.64	59.75
Precision (P %)	Recall (R %)	F-Measure (F %)
Metaphones based synonym mapping (adjectives + other POS categories)		
54.2	**64.3**	58.82
58.4	**60.85**	59.6
59.4	**62.12**	60.73

Table 5. SCL based polarity classification for Urdish data

We rely on two oracles, one that takes care of spelling variations and the other that provides translations. The words that are selected to be translated by the second oracle are carefully chosen based on POS categories that exhibit emotions and pattern switching. We show that the performance of this approach is comparable to what is achieved by training a supervised learning model. In order to identify the POS categories that exhibit pattern switching, we developed a statistical POS tagger for Urdish blog data using a method that does not require annotated data in the target language. Through these two modules (sentiment analysis and POS tagger for Urdish data) we successfully show that the efforts in performing non-topical analysis in Urdu newswire data can easily be extended to work on Urdish data.

8 Future work

Analyzing the test data set for missing and false positives, here are some of the examples of where the model did not work

Example 7.1: *"**tring tring tring tring..** Phone to bar bar bajta hai. **Annoying**." ~ tring tring tring tring tring.. the phone rings repeatedly. Annoying.*

Example 7.2*: "**bookon** ko padna tho ab na mumkin hai. Yaha thak mere friends mujhe **blindee** pukarthey hai" ~ cannot read books any more. Infact, my friends call me blindee.*

Example 7.3: *"Ek Tamana Hai Ke Faqt Mujh Pe Mehrban Raho, Tum Kise Or Ko Dekho To Bura Lagta Hai" ~ I have this one wish that destiny be kind to me If you see someone else I feel bad*

Our method fails to tag sentences like in example 7.1 where English verbs are used by themselves. Our POS tagger fails to capture such stand-alone

verbs as verbs but tags them as nouns. Hence, doesn't occur in the pivot set.

Our second issue is with Morpho syntactic switching, a behavior seen in example 7.2. Nadhkarni (1975) and Pandaripande (1983) have shown that when two or more languages come into contact, there is mutual feature transfer from one language to another. The languages influence each other considerably and constraints associated with free morphemes fail in most cases. The direction and frequency of influence depends on the social status associated with the languages used in mixing. The language that has a high social status tends to use the morphemes of the lower language.

Example 7.4*: Bookon – in books, Fileon – in files, Companiyaa – many companies*

Clearly we can see that English words due to their frequent contact with Urdu grammatical system tend to adopt the morphology associated with the base language and used mostly as native Urdu words. These are some issues, if addressed, will definitely improve the performance of the sentiment analysis model in Urdish data.

References

Abdul-Mageed, M., Diab, M., and Korayem, M. 2011. Subjectivity and Sentiment Analysis of Modern Standard Arabic. *In proceedings of the 49th Meeting of ACL. Portland, Oregon, USA, June 19-24*

Agnihotri, Rama Kant. 1998. Social Psychological Perspectives on Second Language Learning. *Sage Publications, New Delhi*

Bharati, Askhar, Rajeev Sangal and Dipti M Sharma. 2005. Shakti Analyser: SSF Representation

Blitzer, John, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. *In proceedings of the 2006 Conference on EMNLP*, pp. 120–128, Sydney, Australia

Chang, Chih-Chung, Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *In the ACM Transactions on Intelligent Systems and Technology*, Vol 2, no 27, pp 1-27

Dredze, Mark., Blitzer, John., Talukdar, Partha Pratim., Ganchev, Kuzman., Graca, Joao., Pereira, Fernando. 2007. Frustratingly Hard Domain Adaptation for Parsing. *Shared Task of CoNLL*.

Dussias, P. E. 2003. Spanish-English code-mixing at the auxiliary phrase: Evidence from eye-movements. *Revista Internacional de Lingüística Iberoamerican*. Vol 2, pp. 7-34

Gildea, Daniel and Jurafsky, Dan. 2002. Automatic Labeling of Semantic Roles, *Computational Linguistics*, 28(3):245–288

Goyal, P, Manav R. Mital, A. Mukerjee, Achla M. Raina, D. Sharma, P. Shukla, and K Vikram. 2003. Saarthaka - A Bilingual Parser for Hindi, English and code-switching structures. *In proceedings of the 11th Conference of the ECAL*

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, Vol 26, pp. 101–126

Hal Daume III. 2007. Frustratingly easy domain adaptation. *In proceedings of the 45th Meeting of ACL*, pp. 256–263

International Phonetic Association (IPA). 1999. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. *Cambridge: Cambridge University Press*. ISBN 0-521-65236-7 (hb); ISBN 0-521-63751-1

Joshi, Adithya and Bhattacharyya, Pushpak. 2012. Cost and Benefit of Using WordNet Senses for Sentiment Analysis. *LREC*, Istanbul, Turkey

Kachru, Braj. 1978. Conjunct verbs; verbs or verb phrases?. *In proceedings of the XIIth International Congress of Linguistics*. pp. 366-70

Lafferty, John, Andrew McCallum, Pereira. F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In proceedings of the 18th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA . pp. 282–289

Muaz, Ahmed, Aasim Ali, and Sarmad Hussain. 2009. Analysis and Development of Urdu POS Tagged Corpus. *In proceedings of the 7th Workshop on ACL-IJCNLP*, Suntec, Singapore, pp. 24–31, 6-7 August.

Mukund, Smruthi, Rohini K. Srihari. 2010. *A Vector* Space Model for Subjectivity Classification in Urdu aided by Co-Training, *In proceedings of the 23rd COLING*, Beijing, China

Mukund, Smruthi, Debanjan Ghosh, Rohini K. Srihari, 2011. Using Sequence Kernels to Identify Opinion Entities in Urdu. *In Proceedings of CONLL*

Nadkarni, Mangesh. 1975. Bilingualism and Syntactic Change in Konkani Language, vol. 51, pp. 672 C 683.

Pandaripande, R. 1981. Syntax and Semantics of the Passive Construction in selected South Asian Languages. *PhD dissertation. University of Illinois, Illinois*

Prettenhofer, Peter and Benno Stein. 2010. Cross-Lingual Adaptation Using Structural Correspondence Learning. *In proceedings of ACL*

Rasul, Sarwat. 2006. Language Hybridization and Code Mixing in Pakistani Talk Shows. *Bahaudin Zakriya University Journal 2nd Issue.* pp. 29-41

Roark, Brian and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the 2003 Conference of NAACL, HLT - Volume 1* (NAACL '03)

Rie-K. Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *In Jornal of Machine Learning. Res.,* Vol 6, pp. 1817–1853

Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. *University of Pennsylvania, 3rd Revision, 2nd Printing.*

Shimizu, Nobuyuki and Nakagawa, Hiroshi. 2007. Structural Correspondence Learning for Dependency Parsing. *In proceedings of CoNLL Shared Task Session of EMNLP-CoNLL*.

Sinha, R.M.K. and Anil Thakur. 2005. Machine Translation of Bi-lingual Hindi-English (Hinglish) Text. *10th Machine Translation summit (MT Summit X)*

Zentella, Ana Celia. 1997. A bilingual manual on how to raise Spanish Children.