Cross-lingual Sentiment Analysis

Seminar Report

(Spring 2013)

by

Aditya Joshi
Roll No.: 124054003

under the guidance of

Prof. Pushpak Bhattacharyya

Department of Computer Science & Engineering,
Indian Institute of Technology Bombay

Acknowledgement

I thank my advisor, Prof. Pushpak Bhattacharyya for his guidance. The feedback from Prof. Saketha Nath was also extremely useful. I am also grateful to sentiment analysis group (comprising of Ankit Ramteke, Karan Chawla, Raksha Sharma, Vinita Sharma and Nikhil Kumar Jadhav) at IIT Bombay for the weekly discussions which were an enriching experience.

Abstract

Cross-lingual sentiment analysis (SA) deals with predicting opinion expressed in a language (referred to as the *target language*) using resources/approaches developed for another language (referred to as the *source language*). The key challenge is to bridge vocabulary gap between the two languages using cross-lingual projection techniques. One way to do so is to develop a sentiment lexicon in target language using a similar resource available in the source language and a word-linking mechanism between the two languages. On the other hand, ML-based approaches to cross-lingual SA may be categorized as: (A) Relying on translation, (B) Using a common feature sub-space and (C) Other approaches. While translation based approaches require annotated source language data alone, the latter categories of approaches show how labeled data in target language and parallel data between the two language may be beneficial to cross-lingual SA.

Translation using bilingual dictionaries or machine translation systems is a popular approach to cross-lingual SA. Through translation, all documents are mapped to either the source or the target language. However, role of MT in cross-lingual SA has been challenged in two directions: (i) Even if perfect MT is obtained, since translations may not match exactly (for example, words getting mapped to synonyms), perfect cross-lingual SA is not guaranteed, and (ii) Cross-lingual SA using MT performs worse than in-language SA in most cases and in-language SA using an annotated corpus of merely 500 documents is sufficient to surpass cross-lingual SA.

Other approaches to cross-lingual SA map documents of the two languages to a common feature sub-space. We describe approaches including those using (a) word senses (where word sense identifiers form the features), (b) Structural correspondence learning (which uses pivot features to compute projection space), (c) Joint classifier learning (which learns a common feature space for a joint bilingual classifier) and (d) Cross-lingual mixture model (which uses few labeled target language documents). Finally, we describe a different kind of approach based on co-training that uses labeled documents in source and target language.

Contents

List of Figures

Chapter 1

Introduction

The rise of Web 2.0 enabled users to create digital content in the form of blogs, discussion forums, review websites and social networks. This user-generated content on the web contains opinion towards issues, products and people. The value of opinion on the web may be looked at from two perspectives: *business* and *social*. The business of an organization may benefit by tapping opinion expressed by internet users about their products/services. Additionally, businesses may wish to know, in real-time, how their decisions and public announcements have been received by people on the web. The social perspective of online opinion can be understood from the fact that it has mobilized real-world events in the past. Specifically, in recent times, two public movements driven by social media took place in India. These events, namely an anti-corruption demonstration[1] and protests following a sexual assault incident[2], highlight the impact of web opinion on real world events in today's times. The rising impact of digital opinion has accelerated research in sentiment analysis.

This chapter introduces sentiment analysis and cross-lingual sentiment analysis. In section 1.1, we define sentiment analysis (SA) and describe its challenges, approaches and lexical resources. In section 1.2, we introduce cross-lingual SA and present the challenges, approaches and corpora required for cross-lingual SA.

[1]Anna Hazare, Social Media, and the Indian Revolution of 2011 : http://www.domainofhope.com/2011/08/anna-hazare-social-media-and- indian.html

[2]Delhi gangrape: People unite through social media to bring the outrage on streets http://ibnlive.in.com/news/delhi-gangrape-people-unite-through-social-media-to-bring-the-outrage-on-streets/311769-3-244.html

1.1 Sentiment analysis (SA)

Sentiment Analysis (SA) is defined as the task of predicting orientation of opinion in text. The terms *sentiment classification*, *opinion mining* and *opinion prediction* are used interchangeably. The word '*sentiment*' here refers to favourability - the speaker may express favourability/unfavourability towards a target. In SA literature, 'sentiment' is used synonymously with *polarity and opinion*. Alternate definitions of 'sentiment' have been used. Pang and Lee [2007] define sentiment as positive/negative/ objective. Wiebe [1994] define sentiment as a 5-tuple: a tuple of the form *(source, target, feature, polarity, time)*. In other words, sentiment is in the 'polarity' that the 'source' holds with respect to 'target' at a point in 'time'. For example, *'My mother loved the resolution of the camera back in 2002'* would be represented using the tuple as (source = mother, target = camera, feature = resolution, polarity = positive, time = 2002).

Common SA systems, however, assume the simplistic positive/negative/ objective definition of sentiment given by Pang and Lee [2007] and aim to predict that the sentence '*I love butterscotch ice cream*' is positive towards the ice cream; '*I hate the smell of that ice cream*' is negative towards the ice cream and '*The ice cream is made from natural fruit*' is neutral. Thus, SA can be looked as a classification task involving positive, negative or neutral as its output labels. In the sentiment analysis literature, sentiment-bearing sentences (either positive or negative) are called subjective whereas non-sentiment-bearing sentences are called objective. Typically, sentiment analysis research considers one of the following as the target problem:

1. **Subjectivity detection**: Subjectivity detection identifies opinion-bearing portion in text. The output labels for each sentence are subjective/objective. Pang and Lee [2004] states that using subjectivity detection as a first step to sentiment analysis may improve the quality of sentiment analysis. Intuitively, if subjectivity detection is performed before sentiment analysis, a subset of only subjective sentences need to be predicted as positive or negative by the sentiment analysis system.

2. **Positive/negative classification**: Majority of work in SA considers the positive/negative definition. This can be extended to granular positive and negative values ranging over, say, -3 to +3.

3. **SA for different kinds of text**: The nature of input text poses several specific challenges to sentiment prediction. For example, SA of short text like twitter messages involves text normalization but may

often have specific sentiment value. On other hand, for longer documents like product reviews, a writer may talk about different aspects of the product while making a judgment about it. In this case, a SA system is expected to summarize the sentiment expressed. The domain also plays an important role. Text in every domain has typical words which often carry typical sentiment. For example, '*edge of the seat*' is a common term in movie reviews and indicates an exciting aspect of the movie. Similarly, '*read it without putting it down*' in a book review indicates an interesting book.

1.1.1 Challenges

Since sentiment analysis classifies documents as positive and negative, it may be looked at as adapted text classification. However, there are fundamental challenges of sentiment analysis:

1. **Negation**: Negation may be expressed in subtle ways and may change the resultant sentiment. '*The movie is not interesting*' is an elementary form of negation. Negation may also be expressed as '*The movie is not only boring but also poorly made*' (where negation does not act on either phrases that 'but' connects), '*The movie is not boring or poorly made at all.*' (where negation acts on only the first phrase that 'or' connects) or '*The movie is not interesting and innovative*' (where negation acts on both the phrases connected by 'and').

2. **Domain specificity**: Some words have different polarity when used in different domains. For example, the word '*deadly*' in a sports document as in '*He is a deadly football player*' implies positive sentiment. However, in case of a tourist location review as in '*You may have deadly snakes at the camp site at night*' implies negative sentiment. This example has been quoted in Balamurali et al. [2011].

3. **Time dependence**: Time dependence of sentiment can be looked at in two ways: Sentiment held by a word may depend on time and Sentiment may be held with respect to a word at a given point in time. The former case includes words being added progressively to vocabulary of a language or words that imply sentiment implicitly. With respect to the latter, consider the example of a cell phone review. In early cellphones, the ability to send text messages would have implied positive sentiment but in today's times, this features arguably holds little or no value. Another way of looking at sentiment with respect to time is the point in time when a specific sentiment was held by the

speaker. For example, '*I initially hated the camera. But as I started using it, I am now in love with it*' expresses negative sentiment towards the camera in the past but a positive sentence in the present day.

4. **Sarcasm**: Sarcasm involves use of words of a polarity to represent inverse polarity. This makes sarcasm a challenge to SA systems that use polarity of words to determine sentiment. Consider this tweet from *www.twitter.com*: '*Just paid my bills with compliments and nice words, thanks appreciative customers that leave no tip! #sarcasm*'. There are many positive words in this tweet which is, however, negative in itself.

5. **Implicit world knowledge**: Often related to sarcasm are opinion-bearing sentences that may not contain opinion-bearing words at all. In these cases, world knowledge is assumed and helps in conveying the sentiment. Consider the following sentence from *www.cricinfo.com* describing a cricket match. While describing the dismissal of an Indian batsman Jadeja, the speaker states '*Jadeja falls 284 runs short of what would have been a fourth first-class triple-century*'. In fact, implicit knowledge may be required to detect sarcastic statements.

6. **Thwarted expressions**: The phenomenon of thwarting refers to a situation when a subset of expressions thwart or reverse the sentiment expressed by another subset of expressions. The '*expressions*' in this case may be sentences or phrases. The difficulty posed by thwarting lies in the fact that majority-based approaches may not work well. In other words, the overall sentiment in text may not necessarily be sentiment expressed by majority of sentence in it. Consider the example '*The movie may have the finest actors, a talented music director of worldwide acclaim and the most expensive set one has ever seen but it fails to impress*'. In this case, the expression at the end '*..but it fails to impress*' thwarts the positive sentiment expressed in the preceeding sentences.

1.1.2 Approaches

In general, approaches to SA can be broadly classified as:

1. **Lexicon-based approaches**: A lexicon containing sentiment information of words is used to determine sentiment in text. These approaches may use a rule-based classifer to predict the output label.

2. **Corpus-based approaches**: A classifier may be trained on corpus annotated with output labels. Supervised approaches used for SA differ

in aspects including but not limited to feature representation, classifier used, etc.

1.1.3 Lexicons

Sentiment lexicons are resources that contain linguistic units annotated with sentiment information. These linguistic units may be words, phrases or word senses. In addition, sentiment information could be basic output labels, subjectivity information. Some commonly used lexicons for SA are:

1. **SentiWordnet**: SentiWordnet given by Esuli and Sebastiani [2006] is a lexical resource that adds sentiment information to Wordnet in the form of three scores: positive, negative and neutral. The scores in Sentiwordnet have been assigned as a result of an automatic sentiment classifier trained on seed synsets. The sentiment annotation at synset level allows distinguishing between different senses of words that may have different polarity.

2. **WordNet Affect**: Wordnet Affect is a lexical resource that allows determination of affective content of synsets by dividing them into affective categories. Thus, it gives more information as compared to SentiWordNet and is used when analysis to be done is with respect to emotions like anger, joy, etc.

3. **Word lists**: Word-lists like Opinion Finder by Riloff et al. [2005] are also popular. These lists differ in the input (word/sense/phrase, etc.) as well as in terms of the sentiment information. The annotation could be one or more out of 'positive/negative', 'positive/negative/neutral', 'strongly positive/weakly positive', etc.

1.2 Cross-lingual SA

Sentiment analysis of a new language[3] may be performed using one of the two approaches: Developing algorithms, corpora and classifiers for the new language itself, or, leveraging on existing techniques and resources in another language. The latter is called cross-lingual SA. This section provides an introduction to cross-lingual sentiment analysis.

[3]A new language here refers to a language for which existing resources in the form of corpora, lexicons and approaches in the form of classifiers, engineered features do not exist.

1.2.1 Definition

Cross-lingual SA (abbreviated as CLSA) is the task of predicting opinion in documents of language T using documents/resources/approaches developed for language S. The language T and language S are called the target and source language respectively. In the rule-based setting, cross-lingual SA refers to development of lexicons and rule-based classifier for language T using corresponding resources for language S. For the supervised scenario, documents in language S are labeled with polarity and thus, form, the training set for the classifier. Thus, supervised cross-lingual SA involves leveraging polarity-annotated documents in the source language for predicting opinion of documents in the target language.

1.2.2 Challenges

Some challenges of cross-lingual sentiment analysis can be stated as follows:

1. **Reliance on translation**: Since the source and target documents are in different languages, translation is commonly used to obtain both of them in the same language (which may be the source or the target language). However, this often leads to dependence of cross-lingual SA on MT systems - either during training when source language documents are translated to target language or during testing when target language documents are translated to source language. Prettenhofer and Stein [2010] deals with minimizing this error by limiting the number of calls made to MT. In this case, only few words in a document are translated by a weak MT system *i.e.* a system that performs word-to-word translation alone. Balamurali et al. [2013] also chooses to translate source documents into target language before learning the classifier so as to reduce the dependence on the MT system.

2. **Opposite polarity/no polarity**: A common assumption made by cross-lingual approaches is that polarity of words and sentences remains the same across languages. However, there are cases in which this may not be true. For example, in English, the word 'fragile' is used with a negative polarity (with the meaning of being sensitive). However, when translated to Romanian, the word becomes 'fragil' and is used with the meaning of 'unbreakable' and typically does not carry sentiment.

3. **Limited vocabulary**: Assume the case of English-Chinese cross-lingual SA. If a classifier is trained on documents translated from Chinese to English, it is possible that these documents contain words that are not commonly used in natural Chinese sentences.

4. **Instance mismatch due to MT**: Cross-lingual SA depending on MT does not guarantee good performance in case of perfect MT as discussed in Duh et al. [2011]. A sentence may be accurately translated but may have selected synonyms of words which were observed in the training corpus. Such synonyms do not get identified as features and hence, despite meaning the same thing, an instance mismatch occurs.

5. **Colloquial words and idioms**: Idioms specific to languages may not have direct sentiment value if they are not mapped correctly to the other language. For example, the phrase '*foot in mouth*' has a negative polarity in English. However, if this phrase is translated word by word to a different language, it may not have the same polarity. A word by word translation of this phrase to Hindi results in a phrase that is rarely used and does not imply any polarity.

1.2.3 Approaches

A basic technique for cross-lingual SA is **cross-lingual projections**. In the most simplistic case, it may refer to projecting words from source language to the target language or vice-versa. This is a requirement in case of cross-lingual SA since the intersection of words in source and target language can be assumed to be null. The exception here is, however, words that do not have a direct translation in the other language and are used as it is in their transliterated form, if required. Because there are no common words between vocabularies of the two languages, different mechanisms for cross-lingual projections need to be considered. However, it can be noted that translating documents of one of the languages to the other results in documents with possibly the same vocabulary space. Automatic machine translation has been a common approach used for the purpose. This also means that errors in the cross-lingual projection mechanism used also have an impact on performance of cross-lingual SA. In other words, if a document is translated by a weak MT system, the translated version is likely to be poor. It may not preserve all sentiment content of the document and hence, lead cross-lingual SA into misclassifying it. Approaches to cross-lingual SA can be broadly classified as:

1. **Learning the classifier in the source language feature space**: Since the polarity-labeled documents are available in the source language, a classifier may be trained on them as is. The test documents in the target language may then be transferred to the source language for sentiment prediction. The benefit of this approach is that the classifier is trained on natural source language sentences than on translated

sentences which may not be fluent as well as adequate. On the flipside, this approach depends on availability of MT since during testing, each document needs to be translated to be used by the classifier.

2. **Learning the classifier on the target language feature space**: In this approach, documents in the source language are mapped to target language and a classifier is trained on them. This classifier is then used t predict sentiment in test documents. It may be noted that once this classifier has been trained, it does not require availability of MT. It is for this reason that Balamurali et al. [2013] perform their experiments on cross-lingual SA using target language feature space. An alternate approach to learning a classifier in the target language is to adapt an existing classifier for the new language. Brooke et al. [2009] describes how an existing rule-based classifier for English may be adapted for sentiment prediction of Spanish documents.

3. **Learning the classifier in a common feature space**: This category includes approaches that do not project documents to either of the feature spaces of the two languages. Instead, the documents are mapped to a common feature space. In case of Balamurali et al. [2012], this common feature space used is the word sense-based feature space while Lu et al. [2011] learns a joint bilingual classifier for predicting sentiment in both the languages.

1.2.4 Corpora

As mentioned, the supervised setting of a cross-lingual SA system assumes a polarity-annotated dataset consisting of documents in the source language. However, availability of additional corpora can benefit cross-lingual SA. The following additional corpora are considered by some reported research in cross-lingual SA:

1. **Unlabeled corpus in target language**: This type of corpus is used in different approaches, the most noteworthy being the co-training-based approach.Wan [2009]

2. **Labeled corpus in target language**: The size of this dataset is assumed to be much smaller than the training set.

3. **Pseudo-parallel data**: Lu et al. [2011] describe use of pseudo-parallel data for their experiments. Pseudo-parallel data is the set of sentences in the source language that are translated to the target language and

		Target Language(s)	Resources/Tools Required							Approach	
			Labeled corpus (Source lang.)	Labeled corpus (Target lang.)	Parallel corpus	Lexicon (Source language)	MT system	Bilingual dictionary	Unlabeled corpus		Output
Rule-based	Brooke et al [2009]	Spanish				Y				Adapting sentiment classifier from English to Spanish	Lexicon
	Perez-Rosas et al [2012]	Spanish				Y				Using Wordnet linking to create Spanish Sentiwordnet	Lexicon
Using Translation	Mihalcea et al [2007]	Romanian	Y		Y	Y		Y		annotated corpus in target language	Lexicon
	Banea et al [2008]	Romanian	Y				Y			Translating documents to one language using MT	Classifier
	Duh et al [2011]	English, Japanese, French, German	Y				Y			Show that cross-lingual degradation is similar to cross-domain degradation	Classifier
	Balamurali et al [2013]	English, French, German, Russian	Y				Y	Y		Compare different translation approaches with in-language SA	Classifier
Using Common Feature space	Balamurali et al [2012]	Marathi, Hindi	Y				Y			Using sense-based features, achieve cross-lingual SA of resource-scarce languages	Classifier
	Prettenhofer and Stein [2010]	Japanese, French, German							Y	Uses SCL and relies on a weak MT system	Classifier
	Lu et al [2011]	Chinese	Y	Y	Y					parallel corpus and annotated corpora	Classifier
	Meng et al [2012]	Chinese	Y	Y	Y					model	Classifier
Other Approaches	Wan [2009]	Chinese	Y				Y		Y	Unlabeled Chinese documents and labeled English documents co-train two classifiers	Classifier

Figure 1.1: Cross-lingual Approaches: Summary

used as an additional polarity-labeled data set. This allows the classifier to be trained on a larger number of samples.

1.3 Roadmap

Figure 1.1 shows a summary of cross-lingual approaches presented in this report. It may be observed that approaches for European languages have relied on MT while those for languages like Chinese use other cross-lingual projection techniques.

This chapter introduced SA and cross-lingual SA highlighting specific approaches and challenges for both. The rest of the report is organized as follows. Chapter 2 describes rule-based approaches to cross-lingual SA. Chapter 3 onwards describe supervised approaches to cross-lingual SA. Supervised cross-lingual approaches mainly rely on MT for cross-lingual projections. Supervised cross-lingual techniques may be classified as the ones relying purely on MT, the ones using a common feature sub-space and others. Related work presenting benefit of MT and its critique is discussed in Chapter 3. Chapter 4 presents cross-lingual SA techniques that use a common feature sub-space

without transferring documents to either source or target language as such. Chapter 5 describes other approaches to cross-lingual SA. Finally, Chapter 6 presents conclusion and future work.

Chapter 2

Rule-based Approaches

This chapter describes rule-based approaches to cross-lingual SA relying on a sentiment lexicon and a classifier that predicts sentiment in a document using the lexicon. In section 2.1, we discuss how an existing rule-based classifier for a language may be modified for SA of documents in another language. In section 2.2, we describe development of a sentiment lexicon in Spanish using an existing English lexicon. In section 2.3, we describe two systems that involve users by modeling games in order to obtain sentiment-annotated lexicon in a new language.

2.1 Adapting existing rule-based classifiers

This section describes work by Brooke et al. [2009] where the target task is positive/negative classification of Spanish documents. The authors implement a rule-based classifier for sentiment prediction of English documents called Semantic Orientation Calculator (SO-CAL) and show how it may be adapted for Spanish. The classifier uses a dictionary consisting of sentiment words and intensifying words/expressions. The sentiment in a sentence is calculated using polarity of words and manipulated based on: (a) Negation, (b) Intensifying words/expressions and (c) Irrealis markers. The performance of this classifier is evaluated on three corpora: epinions, movie and camera. The performance of the classifier based on which of the additional features are used has been reported. The accuracy of the classifier over the three corpora is 78.7%. If any of the additional rules (*i.e.* negation, intensifying words and irrealis markers) are removed, the accuracy of the classifier goes down. The maximum impact is seen by removing the negation handling feature.

2.1.1 Approach

The authors describe how SO-CAL was modified for Spanish documents. Since Spanish is a highly inflected language, stemmer and tagger are used to obtain root form of words so that they could be looked up in the dictionary. The authors present alternative approaches to create dictionaries for Spanish:

- Existing semantic dictionaries for English can be translated to Spanish. Translations are obtained from www.spanishdict.com and Google Translate.
- Spanish dictionaries already marked with sentiment may be used. These dictionaries are manually corrected wherever the sentiment is observed to be marked incorrectly.
- A Spanish dictionary may be created using a corpus as follows. A set of reviews is downloaded from the internet and the words are marked with parts of speech. The words are then grouped according to their parts of speech and then they are individually annotated with sentiment by manual annotators.
- Dictionaries can also be constructed by combining the above two steps. Such dictionaries are expected to have a higher coverage.

To train the ML-based classifiers, the authors use SVM provided in weka package. It is trained on 2000 documents. There are two sets of classifiers: One that uses English features and another that uses Spanish features.

2.1.2 Evaluation

The evaluation is carried out on English and Spanish corpus from Epinions, Ciao and Doo yoo. The authors show how classifiers (both rule-based and ML-based) trained on Spanish can be used for English and vice-versa. The best accuracy of 76.50% is reported for the English SO-CAL. Among the dictionary-based classifiers, the one created by combining the Spanish sentiment dictionary and dictionary created from Spanish corpus performs the best *i.e.* 71.93%. SVMs perform with an accuracy of around 70-71% which is close to that of the dictionary-based approach.
To understand the effect of translation on cross-lingual SA, the two evaluators, SO-CAL and SVM are compared on original text and translation text. In case of SO-CAL, the accuracy falls from 76.62% yo 71.1% and in case of SVM, it goes down from 72.56% to 69.25%. Thus, expectedly, translation leads to a degradation in the performance. In case of rule-based classifiers, this degradation, however, is more prominent as compared to SVM.

2.2 Creating new lexicons from existing lexicon

Sentiment lexicons are a fundamental resource for cross-lingual sentiment analysis. The most popular sentiment lexicon is SentiWordnet by Esuli and Sebastiani [2006]. Sentiwordnet assigns three polarity scores to each synset of Wordnet: positive, negative and objective. As sentiment analysis research in other languages continues to grow, a need to have a sentiment lexicon in the new languages is being identified. This section describes work by Pérez-Rosas et al. [2012] that presents an approach to building a sentiment and subjectivity lexicon for Spanish. The experiments performed on this lexicon include (a) A sentence sentiment classifier is developed using entries in this lexicon, (b) An approach to incrementally enrich this lexicon is also studied.

2.2.1 Approach

The construction approach leverages on multilinguality of Wordnets of different languages. Wordnets of different languages are connected to each other using synset linking. Thus, corresponding synsets of different languages can be mapped to each other. Using this property, lexicon in a language can be mapped to another language using the synset linking. This work constructs a Spanish sentiment lexicon using SentiWordnet - a sentiment lexical resource for English.

It may be assumed that the quality of the source lexicon determines the quality of the target lexicon. In this context, the strength of a lexicon lies in its coverage as well as accuracy.

To demonstrate how differing these parameters impacts the quality of the lexicon produced, the authors generate two variants of a sentiment lexicon: a full strength lexicon and a medium strength lexicon. The construction of these lexicons can be described as follows:

- **Full strength lexicon**: To build the full strength lexicon, the OpinionFinder list and Sentiwordnet are collectively used as follows. Words marked as strongly positive and strongly negative are selected from the OpinionFinder word list. Since they do not have word sense information, Sentiwordnet is harnessed to select the specific word sense. For this purpose, the word sense corresponding to the maximum polarity that matches with the OpinionFinder polarity is selected. Then, using the synset linking of Wordnets, the word sense is mapped to the Spanish word sense and added to the sentimnt lexicon. This process results in a total of 1347 entries. The lexicon is called a '*full strength*

lexicon' because the OpinionFinder list has manually annotated words. These words are likely to be strong and correct indicators of sentiment. Hence, the '*full strength*' of this lexicon refers to its high precision.

- **Medium strength lexicon**: The medium strength lexicon is built using Sentiwordnet alone. The synset linking of English and Spanish wordnet is used for this projection like in the case of full strength lexicon. Synsets in Sentiwordnet are selected, the synset mapping is looked up and the resultant Spanish word sense is assigned the same polarity label. The resultant lexicon is called a '*medium strength*' lexicon because Sentiwordnet is an automatically created lexicon and likely to have a lower precision than OpinionFinder list.

2.2.2 Evaluation

The sentiment lexicons created for Spanish were evaluated using two gold standard lexicons EsTest1 and EsTest2. Since these lexicons were created for English, additional 100 entries are added by native speakers of Spanish. The evaluation is done using two approaches: SVM trained on feature vectors obtained using INFOMAP software and by manual evaluators. As expected, the positive class F-score of full-strength lexicon is 72.4% while that for medium-strength lexicon is 62.9%. Similar trends are observed for other datasets and negative class as well. This seems to comply with the expectation that the full strength lexicon would have a higher precision than medium strength lexicon. Additionally, the accuracy metrics for SVM are lower than manual evalution as expected. The authors note that the values for negative class are lower than that for the positive class and attribute this to the skewed distribution in the dataset.

2.3 Creating lexicons through crowd-sourcing

In the previous section, we saw an approach to construct lexicon in a language using one in another language. This section describes work by Scharl et al. [2012] and Das and Bandyopadhyay [2010] in which sentiment lexicons for a new language may be created with the help of users through games on crowd-sourcing platforms. Scharl et al. [2012] describes SentimentQuiz, a game developed to obtain annotated words and corpus while Das and Bandyopadhyay [2010] describes Dr. Sentiment, a game to build Sentiwordnets of different languages.

These games prove to be useful on crowd-sourcing platforms like Amazon Mechanical turk.This is useful because because human annotators are quite

unavailable in the requisite numbers for creation of a full-fledged lexicon from scratch.Das and Bandyopadhyay [2010]. The task of annotation can be made interesting using games. Thus, this section highlights how users can be engaged to generate sentiment resources in a new language.

The primary benefit of using games as a medium to obtain annotation is that games reach out to a wide audience of annotators and leverage collective intelligenceScharl et al. [2012]. If engaged sufficiently, these users prove to be intrinsically motivated annotators as they continue to play the game. However, the benefit of these games is not limited to users that play these games over a prolonged duration of time. These games can actually benefit from infrequent users - users who would come across these games while exploring something else and play these games only for a limited time. Since the number of such users is sufficiently large, valuable volume of annotation may be obtained from them.

2.3.1 Design principles

There are multiple avenues where games like these may be hosted. Some of them are listed in Scharl et al. [2012] as follows:

- Internal websites of college portals: The authors report platforms in CMU internal websites
- Crowdsourcing turks: Annotation can be obtained by opening up the task to crowdsourcing platforms. While the users are motivated by small amounts of money, it can be accelerated by embedding annotation into a game.
- Social networking sites: The annotation games could be hosted as applications on social networking websites.Sentiment Quiz, described in Scharl et al. [2012] was hosted on a social networking site.

The authors present certain design ideas implemented in these games and the intuition behind them:

- Reliability of users may be measured by inserting questions with known answers. The confidence in annotation by a user can, thus, be determined based on how many of these known questions the user answers correctly. Scharl et al. [2012]
- Das and Bandyopadhyay [2010] presents an image along with the word to be tagged. The authors state that it helps players visualize the topic being annotated. The image displayed is retrieved from the internet by passing the word as a keyword.

- Scharl et al. [2012] suggests the use of incentives like score boards and game levels. These score boards pitch users against one another. Through a competitive environment, the users can be motivated to annotate more data.

- Scharl et al. [2012] also suggests that judgmets for the same word should be taken from atleast two users. Partial agreements could be aggregated based on user annotation confidence.

Following design principles of sentiment annotation games presented in these papers, we now discuss the two systems.

2.3.2 Sentiment Quiz

Sentiment Quiz described in Scharl et al. [2012] consists of two kinds of annotation: (A) Getting sentences annotated with sentiment, and (B) Getting words annotated with sentiment. (A) leads to a sentiment-annotated corpus whereas (B) leads to a sentiment lexicon. The authors also state that (A) can benefit (B). Words appearing in a large number of documents of specific polarity can be added to the sentiment lexicon being created as a part of (B). The authors used this game with around 3500 users who provide about 325,000 evalations in seven languages, most of which were European. The authors show examples of how sentiment polarity may differ across languages. The following examples are quoted for the English-German pair:

- 'abolish' (-1) versus 'beseitigen' (-0.3)

- 'dirt' (-1) versus 'schmutz' (-0.38)

- 'excitement' (1) versus 'aufregung' (-1)

2.3.3 Dr. Sentiment

Dr. Sentiment proposed in Das and Bandyopadhyay [2010] is a game to create Sentiwordnets in different languages. The game provides support to 56 languages and reports 100 players per day. In Dr. Sentiment, polarized words are obtained from Sentiwordnet and translated using Google translate. The player answers questions pertinent to one of many tasks involving tagging words as +2 to -2. In addition, players are also asked to key in positive and negative words instead of tagging a word. It is possible that through this exercise, several language-specific, culture-specific sentiment-bearing words may come forth.

The authors report manual and automatic evaluation of Sentiwordnets created using the game Dr. Sentiment. Manual evaluation shows culture/age-wise demographics of annotators. The authors observe that the word 'blue' gets tagged negative by users from the Middle East since it is culturally associated with evil. The authors report performance of Sentiwordnet created using this game for Bengali. Using the SentiWordnet and other features like parts of speech, chunks, dependency tree features, etc., the authors achieve 70.04% accuracy.

2.4 Summary

This chapter discusses rule-based approaches to cross-lingual SA. These approaches rely on: use of existing source language lexicon and use of crowdsourcing approaches to generate lexicon in target language. Pérez-Rosas et al. [2012] describe two systems for cross-lingual sentiment analysis of SA: a rule-based approach and a ML-based approach. Some key observations with respect to this work are:

- Different methods to create sentiment dictionaries in target language are presented.
- A rule-based system for cross-lingual SA is compared with a ML-based system. The accuracy figures show that the two perform more or less the same.
- The experiments are performed in both directions *i.e.*, for sentiment analysis of English as well as Spanish.
- The effect of quality of translation is also studied and as expected, it is observed that translation leads to a degradation in performance of sentiment analysis.

Brooke et al. [2009] describe methods to create a sentiment lexicon for Spanish using Wordnet linking. The authors create two lexicons based on expected quality of the source lexicon and present an evaluation of each. We also described how games can be used to engage users in the process of developing a sentiment lexicon. For this purpose, two systems, Sentiment Quiz by Scharl et al. [2012] and Dr. Sentiment by Das and Bandyopadhyay [2010] were described.

Chapter 3

Translation for Cross-lingual SA

A common method for cross-lingual projection is machine translation. In section 3.1, we present one such technique. In section 3.2, we discuss how quality of machine translation may impact its benefit to cross-lingual SA. Finally, section 3.3 presents a critique of machine translation for cross-lingual SA.

3.1 Baseline: Using dictionaries and parallel corpora

Before examining benefit of MT to SA, we show how dictionary and parallel corpora may be used for cross-lingual projection. Towards this goal, this section describes approaches to develop lexicon and annotated corpora in a language using resources available in another language as given by Mihalcea et al. [2007]. In this case, Romanian and English are target and source language respectively. It is believed that an initial process of subjectivity detection improves the quality of sentiment analysis. Hence, the authors focus on developing resources for this task. The authors state that approaches used for developing the lexicon and the corpus are generic and hence, can be mapped to resources for sentiment analysis.

3.1.1 Generation of sentiment lexicon

The first part of Mihalcea et al. [2007] deals with creation of lexicon in target language, Romanian using lexicon in source language, English. As the source lexicon, a subjectivity lexicon of English is used. This lexicon belongs

to OpinionFinder and contains 6856 unique entries. Each entry is marked with its part of speech and a subjectivity label: weak and strong subjective. The lexicon also includes multiword expressions. In order to project sentiment of entries in the English lexicon to Romanian, two English-Romanian dictionaries are used. The first is an authoritative English-Romanian dictionary that consists of 41500 entries while the second dictionary consists of 4500 entries. The Romanian lexicon is, thus, created by looking up an entry in the English lexicon and its entry in the English-Romaniand dictionary. The authors note that since the English lexicon lacks sense information, some errors may be introduced in the resultant lexicon. Also, in order to translate multiword expressions, they are translated word by word and validated using internet search. If internet search returns enough number of results for a given translation of a multiword, it is assumed to be valid and added to the Romanian lexicon. the resultant Romanian lexicon obtained through this process consists of 4983 unique entries including multiword expressions. Like the source lexicon, the target Romanian lexicon annotates words and multiword expressions with part of speech tags and subjectivity labels. The authors observe that using words in their lemmatized form may lead to loss of sentiment - as in the case of an objective word '*memory*' whose plural '*memories*' carries sentiment.

The evaluation of the Romanian lexicon is two-fold: by linguists and by a rule-based classifier. In case of linguistic evaluation, two Romanian speakers annotate 150 entries in the lexicon. These annotators look at the source lexicon entry as well to account for possible incorrect translation or reversal of sentiment. Before annotating any entry as subjective, objective, both or wrong translation, the annotators refer 100 example sentences of each word from the web to ascertain the label. The annotators then discussed their annotations and found out that 123 entries were correct translations. With respect to these correct translations, the interannotator agreement was 0.80 with a Kappa of 0.70. The authors quote examples where a subjective word in English (e.g. fragile) loses its subjective nature when translated to Romanian (e.g. fragil in Romanian refers to breakable objects, without the subjective value).

The lexicon is also evaluated using a rule-based classifier. Sentences are marked as subjective if they contain more than two strong subjective clues. Additional rules for weakly subjective words are also checked to predict subjectivity of sentences. 504 sentences from a Romanian corpus are annotated with subjectivity labels by Romanian annotators. The authors report a precision of 80% and a recall of 20.51% for the rule-based classifier. It is observed that these results are consistent with the reported results of a classifier based on the source English lexicon and evaluated using the MPQA corpus.

3.1.2 Generation of subjectivity-labeled corpus

In the second part of this work, the authors present an approach to obtain sentiment/subjectivity-labeled corpus in Romanian using English as a source language. For this purpose, the authors use a parallel English-Romanian corpus that is aligned at the sentence level. The corpus contains 107 documents and approximately 11000 sentences. The parallel corpus was manually created by a native speaker of Romanian. In order to obtain the desired Romanian corpus, the English portion of the parallel corpus is given as an input to a sentiment classifier in order to obtain their subjectivity labels. This is possible because a classifier for English is available After the English side of the parallel corpus is annotated with subjectivity labels, these labels are mapped to the corresponding Romanian sentences - resulting in a subjectivity-annotated corpus of Romanian sentences. An assumption here is that sentiment is retained across languages *i.e.* a positive sentence in English translates into a positive sentence in Romanian. To validate this, the parallel corpus is manually evaluated by two Romanian annotators who annotate Romanian text with subjectivity labels and an English annotator who annotates English text with subjectivity labels. The Romanian annotators show an agreement of 0.83 while that between English and Romanian annotators is 0.86 - 0.91.

Two versions of the Romanian corpus are generated using two different English document classifiers. The first is a high-preicision classifier while the second is a high-coverage classifier. To evaluate the Romanian subjectivity-labeled corpora created as a result of this process, a Naive Bayes classifier is learned on the corpus.An overall accuracy of about 67.85% is obtained on a gold standard corpus of 504 test sentences. The high-precision classifier has lower overall F-measure than the high-coverage classifier.

3.2 Using Machine Translation

The previous section shows how resources for sentiment analysis in a new language can be developed using bilingual dictionaries and parallel corpus. With existing machine translation systems between language pairs, it is possible to translate sentences in order to project source and target documents to the same language. Banea et al. [2008] study how cross-lingual subjectivity detection relying on pure machine translation (MT) can be achieved.

3.2.1 Approach

The authors present an approach to perform subjectivity detection of Romanian documents using English documents through translation. To this purpose, the authors describe five experiments - four of which examine cross-lingual sentiment analysis using machine translation whereas the fifth experiment gives an upper bound for performance of such a system.

The experiments can be described as follows:

- **Experiment 1**: The English MPQA corpus is translated using a machine translation system from English to Romanian. Speakers of the language annotate the corpus in Romanian with subjectivity. This results in a corpus of Romanian documents annotated with sentiment. A classifier is then learned for these documents.

- **Experiment 2**: The SemCor corpus is manually translated to Romanian. The sentiment labels, however, are obtained through an automatic subjectivity classifier. A classifier is then trained on this corpus obtained.

- **Experiment 3**: The SemCor corpus for Romanian is used for this experiment. Using MT, this corpus is translated to English. Then, a subjectiviy classifier for English predicts subjectivity of the English-translated Romanian documents. The labels are projected back to corresponding documents in the Romanian SemCor Corpus. A Classifier is then trained on these Romanian documents.

- **Experiment 4**: The experiment generates test corpus. A corpus of Romanian sentences is translated to English using a machine translation (MT) system. These English sentences are then passed as an input to a subjectivity detector trained on English documents. These labels are assigned to corresponding Romanian documents. This results in a sentiment-annotated Romanian corpus that is used as the test corpus for the experiments.

- **Experiment 5**: The test corpus in Romanian is translated automatically to English and then its sentiment is predicted. The labeled English documents, thus, obtained form an upper bound performance for the test corpus.

3.2.2 Experiment setup & results

The authors use 504 Romanian sentences as the test corpus and manually annotate them for subjectivity. The MT system used is by Language Weaver

and SVM and Naive Bayes classifiers are trained for subjectivity prediction. The upper bound of F-score is set to 71.83%. It is observed that SVM peforms better than Naive Bayes for all the three experiments.

Experiment 2 that involves manual subjectivity annotation during training set creation performs te best. The authors also report experiments with incremental size of training set and state that at around 40% of the total training corpus, the classifier accuracy began to stagnate.

The authors repeat these experiments for Spanish documents using Google translate Google translate. The authors observe a better overall accuracy in this case. This points to the fact that the quality of machine translation also plays a role in its benefit to subjectivity classification. The lower accuracy for Romanian is also attributed to the fact that Romanian shows more inflections than English.

3.3 Examining influence of MT quality

The assumption of translation based SA is that sentiment remains the same even if a sentence is translated across languages. To evaluate how quality of translation impacts its benefit to cross-lingual SA, Duh et al. [2011] perform experiments on Japanese, French and German as target languages of a subjectivity detection task. This work addresses two key questions:

- **Is the problem of cross-lingual SA solved if perfect MT exists?**: Existing works in cross-lingual SA relate errors in MT to resulting errors in CLSA by stating that poor MT fails to capture sentiment in the source language text. The authors present their findings with respect to the impact of quality of MT on performance of cross-lingual SA.

- **Can other monolingual adaption algorithms be used for cross-lingual adaptations?**: Cross-lingual SA can be viewed as a domain adaptation problem where words in source language S are to be adaped to those in target language T. The authors compare cross-lingual SA with cross-domain SA and present challenges typical to cross-lingual SA.

3.3.1 Approach

Cross-market SA refers to predicting sentiment in documents of a market domain using those of another market domain. The task is non-trivial because there are some words that retain their polarity across markets (for example,

'*excellent*'), there may be words whose polarity gets inverted (for example, '*unpredictable*' is positive if used in movie reviews but negative if used for automobile behaviour). Thus, cross-lingual sentiment analysis can be considered a special case of cross-domain adaptation. This task involves labeled data in source domain S and aims at predicting sentiment for documents in target domain T. In context of cross-market sentiment analysis, the variation in S and T may be as DVD, book, movie, etc. reviews while that for cross-lingual SA is with respect to languages.

The authors perform their experiments on English, Japanese, French and German reviews across three domains Music, DVD and Book. The machine translation system used is Google translate. However, the translation is performed word by word.

3.3.2 Experiment setup & results

This section describes the set of experiments performed and the related observations drawn. The authors report their experiments as follows:

- **Understanding degradation in quality**: The oracle setting corresponds to a situation where the market and the language of source and target data is the same. In other words, no adaptation needs to be performed in this case. This setting forms an upper bound and has an average accuracy of around 80%. The authors then change the language and market in two separate sets of experiments. In cross-market variation, an average degradation of 6% is observed. In case of cross-lingual variation, this degradation is around 7%. Hence, cross-lingual and cross-market tasks appear to be comparably difficult.

- **Revisiting question 1**: The first question that the paper asks is whether problem of cross-lingual SA would be solved if perfect MT existed. The authors, based on their findings in the previous experiment, state that cross-lingual SA will still face additional problems. Instance mismatch is one such issue that they observe. Instance mismatch in test documents may occur in case of unknown words in the classifier feature space. Rephrasing, a word is said to be unknown in the classifier feature space if it occurs in the test document but was not observed in the training document. An example of instance mismatch of this kind is synonyms. Thus, in presence of perfect MT, a sentence may get perfectly translated to the target language - but may use synonyms of words observed in the classifier feature space. Such words may not be detected and hence, not be leveraged by the classifier during prediction.

- **Using transductive SVMs for cross-domain adaptation tasks**: Transductive SVM takes into consideration test samples during training while calculating the parameters. The authors experiment with inductive and transductive SVM for cross-lingual and cross-market tasks. In case of cross-market tasks, adaptive TSVM performs better than inductive SVM in all cases except DVD to MUSIC domain adaptation. This is, however, not the case for cross-lingual adaptation. In many cases (specifically English to German for book and DVD reviews and English to French for Music and DVD revews), inductive SVM performs better than transductive SVM. Based on these findings, the authors conclude that direct application of traditional cross-domain adaptation problems may not be feasible in case of cross-lingual SA. There may be additional intricacies of the problem in case of cross-lingual adaptation. This needs to be incorporated in traditional cross-domain adaptation methods.

3.4 Examining viability of translation

In the previous sections, we discussed how machine translation may be used for cross-lingual SA and how its quality affects cross-lingual prediction. Balamurali et al. [2013] critically examines translation to cross-lingual sentiment analysis vis-a-vis in-language sentiment analysis. Cross-lingual SA relying on different translation approaches are compared to in-language sentiment classification[1].The moot question that is addressed through this work is '*what is better - cross-lingual SA using translated documents or in-language SA using documents/lexicons annotated for the same language?*.'

For the purpose of their experiments, the authors choose to translate the source language documents to target language. This is done to ensure that no additional translation is required throughout the operation of the SA system. The authors also state the difference in effort required for sentiment annotation and collection of parallel corpus for SMT as one of the reasons.

3.4.1 Approach

The authors perform CLSA using different translation approaches as follows:

- **MT-X**: Annotated documents in language S are translated to T using a MT system. A model is learnt on these translated documents. The

[1]In-language sentiment classification corresponds to the setting where the source and target documents belong to the same language

test documents in language T are tested on this classifier. X indicates source language.

- **BD-X**: This experiment represents a situation where MT system does not exist for the given pair of languages. In this cases, a bilingual dictionary is used to translate words one at a time. Annotated documents in language S are translated to T using a bilingual dictionary. A model is learnt on these translated documents. The classifier then predicts the polarity of documents in language T. X indicates source language.

- **MMT-XYZ**: This experiment represents a situation in which documents of multiple source languages are harnessed to learn a classifier for the targer language. This would be a collaborative classifier where multiple resource-rich languages help a resource-scarce language. *XYZ* in the name represents the source languages used. Thus, the experiments are performed as follows. Annotated documents in S_1, S_2... are translated to T using a MT system. A model is learnt on these translated documents. The classifier the predicts the polarity of documents in language T.

- **CoTr-X**: This experiment uses the co-training approach for cross-lingual SA. The co-training approach is described in detail in chapter 5. The advantage is that the classifier can leverage unannotated target language corpus for this process. However, co-training requires two MT systems: one from S to T and another from T to S. This is an added restriction. The co-training approach for source language S and target language T may be summarized as follows:

  - Translate labeled source language document L_S, to target language. Let this corpus be known as L_T'
  - Translate unlabeled larget language documents U_T, to source language. Let this corpus be known as U_S'
  - Repeat the following steps for a pre-determined number of iterations:
    * Learn classifiers on L_S and L_T'
    * From both the classifiers, obtain p positive and n negative most confidently predicted instances
    * Add these instances along with their labels to the training set.
    * Translate these instances and add them along with theirlabels to the training set of the other language

3.4.2 Experiment setup & results

The approaches are compared with in-domain SA for four languages: English, French, German and Russian. The authors downloaded movie reviews from IMDB and selected the highly positive and highly negative reviews based on the user rating specified. Since movie reviews for Russian were not found, book reviews were downloaded. The training corpus size consisted of 3000 positive and 3000 negative reviews for English, German and French and 500 positive and 500 negative reviews for Russian. The Bing translation service was used for translation.

The experiments are reported by increasing the number of training documents from 50 to 400 while the number of test documents remain 200. The experiments were conducted for different combinations of languages. C-SVM available as a part of the LibSVM package was used for training the classifiers.

The authors perform different sets of experiments with each of the four languages as the target language. To explain the notation, consider the following experiments for English as the target language:

- **Self**: A classifier is trained on English documents and tested on English documents. No cross-lingual SA techniques are required since both sets of documents belong to the same language.
- **MT-Fr**: A classifier is trained on translated French dcuments. These documents are translated using MT. **MT-Ru**, **MT-De** represent corresponding experiments using Russian and German documents respectively.
- **BD-Fr**: A classifier is trained on transated French documents. These documents are translated using bilingual dictionaries. **BD-Ru** and **BD-De** represent corresponding experiments using Russian and German documents respectively.
- **MT-FrDe and MT-FrDeRu**: These classifiers are trained on translated documents of French-German and French-German-Russian respectively.
- **Self+MMT-FrDeRu**: The entire set of labeled documents of all languages is considered for training. MT is used wherever required.
- **CoTr-Fr**: Co-training based approach is used for English and French documents. MT systems in both directions are required in this case.

Based on their experiments for different combinations of languages, the authors draw the following observations:

- In all cases, Self forms an upper bound for accuracy. As seen in the case of German documents, the training accuracy of MT-based experiments increases with number of training documents and begins to saturate for about training size = 300 where it is highest at around 80% for Self.
- The Self+MT experiments evaluate if having training data in the target language helps cross-lingual SA. For German as the target language, it is observed that the difference between all approaches is much smaller than in the previous case. However, Self continues to perform the best among other classifiers. This implies that although having labeled data in the target language helps cross-lingual Sentiment analysis, having a classifier trained solely on target language data performs the best.
- The performance of co-training-based approaches shows that Self performs worse than co-traning based approaches for training size of 50. However, beyond a training corpus of 150, the three approaces Co-Tr-En, Co-Tr-Fr and Self perform nearly 80% for German as target language.
- With the constant observation that Self performs the best among all classifiers, the question arises is with respect to generating a labeled corpus for the target language. To understand how much training corpus is sufficient, the accuracy of Self systems is reported for different sizes of training corpus. It is observed that beyond training size of 500 documents, the accuracies begin to saturate at around 500 documents. Thus, although Self needs additional effort in annotation, a good quality sentiment prediction system can be obtained by annotating about 500 documents alone.

Following these findings, the authors conclude that if a sentiment analysis system for a new language is to be created under limited resources, it is worthwhile to spend these resources on obtaining annotated corpus in the same language than relying on a MT system to translate documents from another resource-rich language.

3.5 Summary

This chapter describes cross-lingual SA using translation. Mihalcea et al. [2007] shows how resources in a target language may be created using translation. Salient features of this work are:

- Use of internet search to check if translation of a multi-word expression also forms a valid multi-word expression in the target language

- Linguistic and corpus-based evaluation of the lexicon and the Romanian corpus. The hypothesis made relies on results from both these kinds of evaluation

- The assumption that sentiment is preserved in sentences across languages forms the basis of creation of the Romanian subjectivity-annotated corpus

- Contrary to the popular idea that words retain their sentiment across languages, the authors present examples where this does not occur. Also, the authors show that lemmatization of words may lead to change in the subjectivity label as in the case of the word '*memories*'.

Banea et al. [2008] describe how MT can be used for training sentiment classifiers either in the source of target language. The pecularities of this work are as follows:

- The paper compares automatic and manual translation, and automatic and manual subjectivity annotation.

- The experiments for Spanish documents perform with a higher overall accuracy. This points out that quality of machine translation also determines its benefit to cross-lingual sentiment analysis.

Following dependence on MT as a cross-lingual projection technique, Duh et al. [2011] validate how quality of MT impacts its benefit to cross-lingual SA. Key arguments made in this paper are:

- Cross-lingual adaptation may be looked at from the perspective of a general cross-domain adaptation problem.

- The experiments use reviews from four languages: Englis, Japanese, German and French and three domains: Book, Music and DVD. With the help of these experiments, the authors establish comparisons between the two problems in question.

- The authors note that although translation may be correct, there may be instance mismatch as in case of synonyms appearing in the translated versions.

- By showing that cross-domain adaptation techniques do not necessarily perform better than general techniques for cross-lingual SA, the authors point towards improved adaptation algorithms for cross-lingual SA.

Finally, Balamurali et al. [2013] evaluate the viability of translation for cross-lingual SA. The authors compare in-language SA with cross-lingual SA and show that:

- In-language SA consistently performs better than cross-lingual SA. Hence, in the light of limited resources, investment in developing tools and resources for the target languages should be a priority over using MT systems for cross-lingual SA.

- The authors also present different ways in which cross-lingual SA using MT and bilingual dictionaries can be performed. One of these experiments, specifically MMT, involves use of multiple source languages to train a classifier for the target language.

- The experiments show that using documents from the target language along with those from other language also benefits cross-lingual sentiment analysis. In other words, cross-lingual SA benefits if labeled documents from the target language are available.

- The authors also present an optimal number of documents in the target language that give a sufficiently good accuracy for in-language SA. The fact that only small number of documents are sufficient in this case highlights the value of in-language SA.

- The authors also point out the intuition behind selection of languages for experimentation. These languages are politically and commercially important languages. Hence, they have SMT systems of better quality than many other widely spoken languages. The problems using MT demonstrated in this work may be more severe for these other languages where good quality MT does not exist.

Chapter 4

Common Feature Sub-space for Cross-lingual SA

The previous chapter described how MT may be used for cross-lingual SA. MT maps documents to the feature space of either of the languages - based on the language in which the classifier is trained. However, it is possible to consider a feature sub-space common to both the languages. In this chapter, we discuss different feature sub-spaces and approaches to map source and target language documents to this common sub-space to learn a classifier. In section 4.1, we discuss how word sense-based feature space can be used. Section 4.2 shows how cross-lingual SA may be modeled a cross-domain adaptation task and structural correspondence learning be used. In section 4.3, cross-lingual SA using a joint bilingual classifier is discussed. Finally, section 4.4 describes a cross-lingual mixture model for cross-lingual sentiment classification. It must be noted that each of these works do not map documents to either source or target document words but to a set of hyper-features common to both the languages.

4.1 Using sense-based feature space

Majority of cross-lingual sentiment analysis approaches define the source language as a resource-rich language while the target language as a resource-scarce language. Balamurali et al. [2012] digress from this definition and look at how two Indian languages, Hindi and Marathi - none of them particularly resource-rich, can be used for cross-lingual SA. The requirement for using the sense-space-based approach reported in this work is a common sense space based on linked Wordnets.

This work is based on Balamurali et al. [2011] which shows that word

senses prove to be better features than words alone. Hence, words in labeled corpus are annotated with synset identifiers from Wordnet. These identifiers are used as features for the classifiers. In addition to improved performance, the use of sense space based on synset identifiers has the following benefits:

- The sense space allows representation of documents using lower number of features.
- The word space is unable to capture words in the test corpus which do not occur in the training corpus but whose synonyms do. On the other hand, the sense space replaces both these words with the same synset identifier thereby recognizing the commonality between them.

4.1.1 Approach

The approaches studied in this work are: (a) **In-language SA**: This is the monolingual setting for SA *i.e.* the source and target language are the same. This system gives the skyline performance of SA for the two languages. (b) **Naive translation**: Since a pair of MT for Hindi and Marathi does not exist, the authors simulate a naive translation system based on lexeme replacement. This MT system is compared with the sense-based approach. Two variants of the naive translation system are presented. (c) **Sense-based approach**: A classifier is trained on documents of source language using word senses as features. The authors show how this classifier can be used to perform cross-lingual sentiment analysis.

In-language SA

In-language SA refers to the mono-lingual setting of sentiment analysis. In this case, the source language is same as the target language. This set of experiments aim to establish benefit of word senses as features for sentiment analysis of Hindi and Marathi. Classifiers with following feature representations are trained for sentiment prediction: (a) Words (W), (b) Human-annotated senses (M), (c) Senses annotated by automatic WSD (I), (d) Words and human-annotated senses, (W+S(M)) and (e) Words and senses annotated by automatic WSD (W+S(I)). The authors state that the classifiers (d) and (e) arise from the fact that Hindi and English wordnets are incomplete and hence, many words may not have any synset identifiers. Hence, if the feature space is limited to word senses, the contribution of these words to the classifier may be lost.

Naive translation

The authors simulate a naive translation system for Marathi and Hindi since there exists none between the two. For this purpose, Multidict, a manually created linking dictionary is used. Multidict consists of two aligments: synset-level alignments and word-level alignments. Corresponding synsets of the two languages are linked using synset-level alignments. The word-level alignment operates within a synset. Specific words within a synset of a language are linked with specific words within a synset of another language. Thus, word-level alignments provide a more accurate translation of a word. This linking is referred to as 'cross-linking of words and synsets'.

Using multi-dict, the authors implement two naive strategies of translation:

- Exact replacement (E): Using the word and the disambiguated synset identifier, the exact word is located and selected as the translation.
- Random replacement (R): Uisng the word and the disambiguated synset identifier, any random word from the synset of the target language is selected.

The exact replacement technique gives better translation than random replacement. Clearly, the translated sentence may not have grammatical structure or inflected forms of words.

Sense-based approach

This is the approach that the authors introduce in this work. The documents in different languages are represented in feature space consisting of word senses. Because of the word sense representation, classifier trained on one language can be used for sentiment prediction of documents in another language. Thus, the process works as follows. The words in labeled documents are disambiguated by either manual annotators or automatic WSD techniques. A classifier is then learnt using an instance defined by sense identifiers. The words of a test document are annotated with senses before being predicted using this classifier. The authors present experiments by varying the following:

- **Target language**: Hindi, Marathi
- **Feature representation**: As in the case of in-language SA, features used may be: manually annotated senses (M), senses annotated by

automatic WSD (I), words and manually annotated senses (W+S(M)) and words and automatically annotated senses (W+S(I)).

4.1.2 Experiment setup & results

The datasets for Hindi and English consist of travel destination reviews. The Hindi corpus consists of 100 positive and 100 negative reviews while the Marathi corpus consists of 75 positive and 75 negative reviews. The reviews are manually labeled with sentiment as positive or negative. Two versions of this labeled corpora are creation: one in which words are manually labeled with senses and another in which words are labeled with senses using IWSD, an automatic word sense disambiguation engine. C-SVM provided as a part of LibSVM package is used as the classifiers.

The experiments and their results are as follows:

- **In-language SA** In-language SA evaluates how beneficial word sense-based features are to sentiment classification of Hindi and Marathi. As expected, for Hindi as well as Marathi, W+S(M) classifier performs with the highest accuracy (83.06% in case of Hindi and 97.87% in case of Marathi). The overall accuracies for Hindi are lower than that for Marathi. This may be attributed to the relatively smaller size of the Marathi corpus and also to the fact that Marathi Wordnet consists of fewer senses. In case of Marathi, classifier trained on words performs with an accuracy for 86.3%. An improvement of around 7% is seen in case of classifiers trained on automatically annotated senses while it is around 11% in case of classifiers traind on manually annotated senses. It seems that manually annotated senses have better precision and hence, are more accurate with respect to sentiment in documents.

- **Naive translation v/s sense-based approach for cross-lingual SA**

  These experiments deal with cross-lingual setting of SA *i.e.* where source language is different from target language. Two sets of experiments, one for Hindi as the target language and one for Marathi as the target language are carried out. These experiments compare cross-lingual SA using naive translation as described above with cross-lingual SA using sense space.

  For cross-lingual sentiment analysis for Marathi documents, using W(E) *i.e.* exact replacement gives an accuracy of 71.64% while that for W(R) *i.e.* random replacement is 70.15%. This is in accordance with the intuition that random replacement may not always select accurate word

for translation. The sense-based representation performs better than SA using translation byabove 14%. Both manually annotated senses and automatically annotated senses exhibit similar performance. For cross-lingual sentiment analysis for Hindi documents, manually annotated senses (M) perform with the highest accuracy of 72.08%. In this case, the difference between M and I is of about 4%.

The authors attribute the lower accuracy for cross-lingual SA of Hindi documents to the fact that Marathi Wordnet is generated from Hindi Wornet. Hence, Marathi Wordnet may not have all concepts present in Hindi Wordnet leading to a lower coverage of features. Additionally, the authors also point out a defect in the Hindi morphological analyzer where candidate senses of verbs could not be matched due to inflection. This leads to lower precision of sense annotation in case of Hindi documents.

4.2 Using structural correspondence learning (SCL)

An alternate approach may use structural correspondence learning, a technique for cross-domain adaptation. Prettenhofer and Stein [2010] show that cross-lingual SA may be considered a cross-domain adaptation task where documents of domain S (in this case, language S) are used for sentiment prediction of documents of domain T (in this case, language T). Formally, a cross-lingual SA setting corresponds to: A classifier is learned on source language S using parameters from vocabulary V_S, the vocabulary of the source language. This classifier is used to predict documents in language T where a document is made up of words from vocabulary V_T. The core issue with cross-lingual sentiment analysis is that $V_S \cap V_T = \emptyset$. Thus, the challenge to a cross-lingual SA system is to how to map V_S to V_T and/or vice-versa. The authors experiment with Japanese, French and German documents.

The structural correspondence learning framework has the following requirements with respect to the corpora:

- Labeled corpus in the source language: The sentiment labels of words are learned from this corpus.
- Unlabeled corpus in target language: The corpus helps the classifier learn the kind of patterns to expect in test data. As will be evident, this corpus contributes to selection of pivot features and calculation of correlation between them and other words in the corpus.

- No parallel corpus: This approach does not require parallel corpus in the two languages. This is an added advantage since procuring a sizeable parallel corpus for two language may be a costly task.
- Unlabeled corpus in source language: This is an optional corpus. It adds similar value as the unlabeled corpus in target language.
- A weak MT system: The approach uses a weak MT system - a system that performs word-to-word translation only and assumes that one word in source language is translated to exactly one word in the target language. This work differs from the rest that use MT for cross-lingual projections in that it restricts the number of calls to a MT system by fixing a budget *i.e.* maximum number of calls that can be made.

The source language corpora D_S and D_S, UT contain words in V_S and no words in V_T while the target language corpus D_T, U cotains words in V_T but no words in V_S. The sentiment labels are shown on far right. It can be seen that only D_S possesses positive and negative sentiment labels.

4.2.1 Overview of SCL

Structural Correspondence learning (SCL) aims to identify common substructures across domains in order to perform cross-domain adaptation. SCL relies on key features common to both domains *i.e.* features that behave in a similar manner with respect to the output in both the domains. These features are referred to as pivots. A desirable property of pivots is that they should have high support *i.e.*, they should be discriminative with respect to the output and high confidence *i.e.*, they should have sufficient occurrence in the corpus. SCL selects such pivot features and learns linear predictors to predict these pivot features using other words in the document. Then, correlation between these pivot features is calculated resulting in a set of projection parameters. Thus, a classifier can be trained by projecting documents of different domains to this new space of projected parameters. Thus, this classifier helps in prediction of both the domains.

4.2.2 Approach

The steps to perform cross-lingual SA using SCL are summarized as follows:

- **Determination of pivots**: A pivot in case of cross-lingual SA is a pair of words w_s, w_t which are translations of each other and w_s occurs in V_S while w_t occurs in V_T. Pivot features are strong indicators of desired

output labels. An example of a pivot for English-Hindi cross-lingual SA is excellent, behtareen where behtareen is a translation of excellent in Hindi. This is a good pivot because: (a) excellent, behtareen are strong indicators of positive sentiment in their respective languages (this corresponds to their confidence as a pivot), and (b) They occur frequently in documents of their respective languages (this corresponds to their support as a pivot). Hence, selection of pivots is performed as follows:

  - Find top V_P words that have highest mutual information with the class label. Here, the number of V_P words selected is much less than V_S, the total vocabulary of the source language. Since they are strong indicators of sentiment, we expect that polarized words get correctly captured. In other words, a high confidence can be assured for these words that have been selected. It must be noted that the mutual information with class label can be calculated only from the source corpus. Hence, the superset of pivot features is derived only from the source language vocabulary.
  - Using a translation system, words selected in the previous step are translated to obtain a wt for every ws. A w_s, w_t pair then forms a pivot.
  - To ensure that there is adequate support for pivots and only such pivots get selected, candidate elimination is performed. w_s,w_t that occur more than a minimum threshold of times in the corpora are retained. This information can be calculated from source as well as target language unlabeled and labeled corpora.

- The above steps result in a set of m pivot features w_s, w_t.

  - **Predictors for pivots**: The second step is to find predictors for pivots using other words. This is same as predicting whether a pivot occurs in a document given all other words in the document. To learn these predictors, a dataset consisting of MASK (x,pl) as featuers and IN(x,pl) is created for the unlabeled corpora. MASK(x,pl) returns a vector of all words in the document other than the pivot words. IN(x,pl) is an indicator function that returns whether or not the given pivot word is present in the corpus. Parameters for these linear predictors are learned.
  - **Finding projection parameters**: The goal of this step is to find underlying structures common to both the languages. To do so, singular valued decomposition (SVD) is performed on the weight

vector obtained as a result of the previous step. The result of SVD is the matrix U. Top k values in U are returned as projection parameters. These are the best parameters to map non-pivot features to pivot-features.

– **Classifier training**: The classifier for cross-lingual sentiment analysis is learnt using these projected values. The classifier is trained using the regularized loss minimization function.

4.2.3 Experiment setup & results

The datasets belong to three domains: book, DVD and music. Reviews of these domains are downloaded and the ones with user rating greater than +3 or less than -3 are selected. The selection on the basis of rating is to ensure that strongly opinion-bearing reviews are selected. The resultant is a training corpus of 2000 documents, a test corpus of 2000 documents and an unlabeled corpus of 9000-50000 documents per domain, per language.

The baseline experiment consist of a classifier trained on source language documents. The test documents in the target language are translated into the source language and predicted using this classifier. This experiment is shown as CL-MT in the table. The upper bound consists of prediction using an in-domain/in-language classifier. The approach reported in this work is represented as CL-SCL.

CL-MT performs better than CL-SCL in many cases (the most significant being Japanese book reviews where CL-MT performs with an accuracy of 70.22% while that for CL-SCL is 73.09%) though the difference with respect to the upper bound is comparable in case of CL-SCL and CL-MT. For German documents, CL-MT and CL-SCL both perform with a mean accuracy of about 77-79% . The results are similar for othe cases except in the case of French music reviews and Japanese music reviews. In these cases, CL-MT differs by 10.31% from the upper bound. On the other hand, CL-SCL shows a deviation of 7-8% from the upper bound. Thus, the table shows that CL-SCL performs close enough to CL-MT.

The authors also examine how performance differs with number of pivots selected. CL-MT performs better thanCL-SCL until a value of m approximately equal to 450. Beyond that, CL-SCL imitates the performance of CL-MT. From these two findings, it is seen that CL-SCL does not surpass CL-MT performance in most cases. However, the value of CL-SCL lies in the fact that it uses substantially low number of machine translation calls and also makes use of unlabeled corpora in both the languages.

4.3 Joint bilingual classifier learning

Lu et al. [2011] present an alternative approach of performing sentiment analysis of Chinese documents. The demarcation between languages as resource-rich and resource-scarce is not seen in this case. Instead, the authors propose a model where a sentiment classifier for both the languages is jointly learned. This work relies on sentiment labeled corpus in the two languages and an unlabeled parallel corpus. The objective of this work is to maximize the regularized joint likelihood of language-specific data which is obtained through the labeled data and inferred sentiment labels obtained as a result of unlabeled parallel text.

4.3.1 Approach

The labeled corpus is indicated by D_1 and D_2 while the unlabeled corpus is indicated by U. The unlabeled corpus U consists of (X_1',X_2') were X_1' is from language L_1 and X_2' is from language L_2. On the other hand, the labeled data for language L_1 is (X_1, Y_1) while that for language L_2 is (X_2, Y_2).

Two sets of parameters θ_1 and θ_2 are learnt for the two languages. The likelihood of θ_1 and θ_2 given the unlabeled parallel data and labeled data in the two languages is given as:
$L(\theta_1, \theta_2|D_1, D_2, U) = p(Y_1|X_1; \theta_1)p(Y_2|X_2; \theta_2)p(Y_1', Y_2'|X_1', X_2'; \theta_1, \theta_2)$. The first term corresponds to probability of labels for documents in D1 and D2. The second term corresponds to probability of labels of parallel documents being the same. The third term is the regularization constant.

The EM algorithm can be described as follows:

- **Step I: Initialization**: Based on the labeled data, two initial monolingual models are trained. The parameters learnt are iteratively modified using the EM algorithm until convergence.
- **E-step**: The value of P(y—x) is computed for each example in the three corpora. The expectation of log likelihood is calculated.
- **M-step**: The new values of the parameters are determined by maximizing the regularized joint log likelihood.
- The algorithm halts if the difference in parameter values is small.

4.3.2 Experiment setup & results

The labeled corpus for English consists of MPQA corpus in which about 5000 sentences were labeled with subjectivity for the purpose of question answering. The labeled English corpu is NTCIR-EN and consists of about 1737 news text sentences labeled with setiment. The labeled Chinese corpus consists of 4294 sentences from the NTCIR-CH corpus. The unlabeled parallel corpus is the ISI English corpus consisting of 20000 parallel sentences. In addition to these corpora, the authors also experiment with pseudo-parallel data - data generated by obtaining translations using Google Translate.

Five-fold cross-validation experiments are conducted using SVM, MaxEnt, Monolingual T-SVM, bilingual T-SVM and Co-training SVM. The experimental results are presented for two settings: One with NTCIR-EN as training data and NTCIR-CH as test data and another with MPQA as training data an NTCIR-CH as test data. The Joint-model based method performs with an accuracy of 79.29% for English and 83.54% for Chinese which is better than the baseline experiments in all cases. Among the baselines, Co-SVM performs the best with an accuracy of 82.63% for the corresponding Chinese setting.

The authors perform an experiment to study change in accuracy depending on weight and size of unlabeled data. For increasing values of unlabeled data weight, the accuracy of four systems is presented: English and Chinese accuracy for the two settings of training corpora. When the weight assigned to unlabeled data is zero *i.e.* unlabeled data is not incorporated while training the classifier. As the weight of unlabeled data goes on increasing, the accuracy improves significantly. Beyond the weight value of 0.4, the accuracy stabilizes for all four settings. Also, as the size of unlabeled data, the accuracy of the four systems improves significantly. However, beyond 2000 unlabeled sentences, the performance saturates. This means that with about 2000 unlabeled sentences, substantial improvement in accuracy of sentiment classification is obtained for both the languages for whom the classifier has been jointly learned. The authors also present their findings with respect to use of pseudo-parallel data. With addition of pseudo-parallel data CH-¿EN, the classifier for English sees a significant improvement. This is expected since thi means addition of more English examples.

The authors observe that the classifiers agree on 73.87% predictions in case they are not learnt jointly. However, if they are learnt jointly, these classifiers agree on 99.89%. This agreement is intuitive since parallel sentences are assumed to have the same sentiment label. The authors also present features with maximum weight change. Among the positive class features in this list are 'important'and 'cooperation' while among the negative ones are

'difficulty' and 'not'.

4.4 Using cross-lingual mixture model

Meng et al. [2012] present a mixture model for cross-lingual sentiment classification of Chinese documents. The authors state that cross-lingual SA using MT introduces typical errors. The sources of these errors can be broadly classified as: (a) **Errors in MT**: The fundamental assumption of cross-lingual sentiment analysis is that a sentence retains its polarity across languages. This means that the sentiment in a sentence written in a language is same as its translation into another language. An error in MT could possibly mean that meaning/semantics of the sentence are not correctly transferred. This impacts the quality of a sentiment analysis system using cross-lingual adaptation. (b)**Limited vocabulary**: The vocabulary learned by a classifier is the vocabulary of labeled data set used for training.

The model takes into consideration sentiment indicators in both source and taget language and attempts to uncover hidden associations between these words of the two languages. The model being proposed is based on the generative framework and assumes that words in source as well as target language documents are generated as a result of an underlying distribution of words. The goal of the cross-lingual mixture model is, thus, to maximize likelihood of bilingual parallel data with respect to the output labels. As a result, the model uncovers several related words across the two languages. Since the parameters are learned over source as well as target language features, the cross-lingual mixture model is able to reduce the impact of limited coverage of the vocabulary.

The cross-lingual mixture model aims to synchronize generation of words in source and target language by making the generative assumption. By facilitating an improvement in vocabulary coverage, the algorithm is able to transfer polarity label information between source and target language. Thus, the task here is to predict sentiment in documents of language T using documents of language S. This approach takes as input a labeled corpus in source language, unlabeled parallel corpus and labeled corpus in target language. The labeled corpus in target language is optional. In absence of labeled target language data, the accuracy of the system is around 71% while with the labeled target language data, it improves substantially to 83%.

The intuition behind the mixture model can be explained as follows. A polar Chinese sentence can be generated in two ways. One way to do so is to directly generate Chinese words according to the polarity of the sentence. Alternately, a Chinese sentence may be generated by first generating English

words of the desired polarity and then translating them to Chinese. This cross-lingual mixture-based approach incorporates both these cases in the parameters of its model allowing a mixed representation of words across the two languages. The hidden variables in this case are defined as the polarities while the words in the corpus are observed.

4.4.1 Approach

The cross-lingual mixture model deals with three types of probabilities:

- **Document class generation**: This is the probability of generating a polarity class C_S from a Bernoulli distribution. The document class generation probability is given as PS(C). The candidate polarity classes in this case are positive and negative.

- **Word generation**: This probabiliy corresponding to the first part of generating words of a language corresponding to a given polarity as given above. This probability is defined as the probability of generating source language words w_S from a multinomial distribution. This probability is given by P(w_S — C_S).

- **Word projection**: The second part of the generative model was defined as the process of generating words of a given polarity in a language and translating them. The cross-lingual mixture model incorporates this using word projection probability. It is the probability of projecting the source language words w_S to target language words w_T. In other words, this probability represents the chance that word w_T is generated from w_S.

The likelihood parameters for the three corpora are defined as follows:

- **Unlabeled parallel data**: The likelihood of parameters for unlabeled parallel data can be given as:
$$L(\theta|U) = \sum_{i=1}^{|U_s|}\sum_{j=1}^{C}\sum_{s=1}^{|V_s|}[N_{si}log(P(w_s|c_j) + P(w_s|w_t)P(w_t|c_j))]$$
$$+ \sum_{i=1}^{|U_t|}\sum_{j=1}^{C}\sum_{t=1}^{|V_t|}[N_{ti}log(P(w_t|c_j) + P(w_t|w_s)P(w_s|c_j))]$$
U_s and U_t indicate the number of unlabeled source and targte documents. —C— indicates the number of category laels and V_s and V_t are the source and target vocabulary respectively. There are two additive components to the formula: the first generates the words in the source language and the second generates the words in the target language.

Consider the first component. Here, N_{si} is the number of occurrences of word s in document i. The first component can be read as the probability of generating words in the source language by generating them directly from a given class label or generating a word in the target language and then projecting it to the source language. Similarly, the second component gives the probability for generation of target language words where N_{ti} is the number of occurrences of words wt in document di.

- Labeled source data: The likelihood of parameters for the source data is given as:
  $L(\theta|D_s) = \sum_{i=1}^{|D_s|} \sum_{j=1}^{|C|} \sum_{s=1}^{|V_s|} N_{si} log P(w_s|c_j) \delta_{ij}$
  In this case, δ_{ij} is an indicator function. It is fired for i = k and j = c if the category of document k is c. Thus, this likelihood deals with word generation on the source language side.

- Labeled target data: The likelihood of parameters for the target data is given as:
  $L(\theta|D_t) = \sum_{i=1}^{|D_t|} \sum_{j=1}^{|C|} \sum_{t=1}^{|V_t|} N_{ti} log P(w_t|c_j) \delta_{ij}$ In this case as well, δ_{ij} is an indicator function that is fired for i = k and j = c if the category of document k is c.

To estimate values of parameters, an EM algorithm is run that iteratively tunes these values. The EM algorithm is described as follows:

- **E-step**: The E-step calculates probability of output label given the document of the corpus by using the word generation probabilities for source and target language. $c_u si$ denotes the category of ith document in the unlabeled source language corpus while $c_u ti$ denotes the category of the ith document in the unlabeled target lanuage corpus.

- **M-step**: The M-step calculates the word generation probability for all words in the vocabulary for source and target languages, both. The numerator shows that for all documents in the source corpus, the probability of category given the document is used. Add-one smoothing is used to accommodate unknown words.

The EM algorithm iterates until convergence to return the expected probability values based on observed data.

4.4.2 Experiment setup & Results

The unlabeled parallel data is from ISI parallel corpus. The Chinese labeled corpus is from NTCIR-CH while the English labeled corpora used are MPQA and NTCIR-EN. The background experiments that are compared with CLMM are: MT-SVM (SVM based on machine translation), SVM (pure SVM), MT-Co-train (Co-training approach), Para-CoTrain (co-training approach using an additional set of unlabeled English sentences) and Joint-Train (maximum entropy classifiers over both the languages)

The first experiment deals with using only English labeled data. There are two settings that are explored: (a) NTCIR-EN as the training corpus and NTCIR-CH as the test corpus and (b) MPQA as the training corpus and NTCIR-CH as the test corpus. No labeled Chinese documents are used in this experiment. CLMM performs the best among MT-SVM, MT-Cotrain and Para-Cotrain for both the training corpus settings.

The second experiment aims to understand the benefit of using labeled data in target language. Thus, for this set of experiments, the authrs use labeled data for English as well as Chinese. Five-fold cross-validation is performed. In case of NTCIR-EN as training corpus, Joint-Train performs better than CLMM wherein the accuracy of CLMM is 82.73%. In case of MPQA as training corpus as well, CLMM has an accuracy of 83.02% while Joint-Train performs marginally better. However, the benefit of CLMM is shown through comparison of running times of the thee experiments. Para-Cotrain requires 100 iterations and takes about 6 hours. Joint-Train takes 55 seconds for 10 iterations. CLMM requires 30 seconds for 10 iterations which is much faster than Para-Cotrain and half as much time as Joint-Train.

The third experiment studies the impact of size of parallel data on the performance of CLMM. The diagram shows performance for the setting where MPQA is used as the training corpus. The difference between CLMM and MT-Cotrain and Para-train is negligible for small training sizes. However, the differences between CLMM and other approaches increase sharply with the training corpus size. CLMM achieves saturation at abut 12000 sentences ith an accuracy of around 70% wherein MT-Cotrain and Para-Cotrain perform with n accuracy of 57-58%.

The fourth experiment studies the impact of size of labeled target language data on the performance of CLMM. For about 500 labeled sentences, CLMM does slightly better than Para-CoTrain and Joint-Train. These approaches are significantly better than vanilla SVM which performs with an accuracy lower than 65%. As the number of labeled sentences increases, the performance of Para-Cotrain, Join-Train and CLMM improves. At labeled data size of 3500 sentences, the three approaches converge to an accuracy

more than 80%. Thus, as the size of labeled data increases, it is observed that the three methods exhibit comparable performance.

4.5 Summary

This chapter described four approaches to cross-lingual SA using a common feature subspace. First of all, Balamurali et al. [2012] present an approach based onWordnet synsets. The pecularities of this work are:

- This work digresses from the traditional resource-rich/resource-scarce definition of languages in case of cross-lingual adaptation tasks and looks at it as one language aiding sentiment classification of another.
- In absence of a MT system, the authors show two methods of leveraging synset linkings available for the two languages. While exact replacement replaces a word with its exact translation, random replacement replaces a word with any word in the linked synset.
- The results show that sense-based approach performs better than translation-based approach for target language Hindi as well as Marathi.

Prettenhofer and Stein [2010] present a cross-lingual SA technique modeled as a cross-domain adaptation task. A SCL-based approach is used to derive pivot features which are strong indicators of sentiment. Some key arguments of this work are:

- Determination of pivot features. While the confidence is determined from the source corpus, the support is derived from both the source and the target language corpus.
- SCL allows representation of underlying structures between features of both the languages by finding projections of parameters .
- Though CL-SCL performs only as good as CL-MT, it scores on the count that it has a lower reliance on a MT system. In addition, only a weak MT system that performs word-to-word translation is sufficient for CL-SCL to perform well.

Lu et al. [2011] present a method to jointly learn classifiers for the source and target language. The features of this experiment are:

- The model incorporates probability of labels of parallel documents in unlabeled parallel data being the same. This relaxes the commonly believed assumption that sentences across languages have the same polarity label.

- The authors validate how pseudo-parallel unlabeled data helps joint learning of sentiment classifiers.
- The authors justify the value of their approach by observing increased classifier agreement and feature weight changes for strongly positive and strongly negative words.

Meng et al. [2012] present a cross-lingual mixture model for sentimtent classifcation. In this work, the authors show that:

- A model based on the generative framework is able to reduce the impact of limited vocabulary in case of cross-lingual sentiment classification.
- Such a model assumes that a word of language L may be generated in two ways: Generated from a given polarity class directly or generated from a given polarity class in another language and then projected to language L.
- The authors show that CLMM performs as well as co-training-based approaches but takes significantly less time.

Chapter 5

Other ML-Based Approaches

This chapter describes an approach to cross-lingual SA that performs co-training of classifiers for the two languages.

5.1 Using co-training

The vanilla approaches to cross-lingual sentiment classification assume that the source language has a sentiment-annotated corpus or a sentiment-annotated lexicon or a sentiment classifier. There is no assumption made for target language. Wan [2009] describes a co-training-based approach to perform sentiment classification of Chinese documents as positive and negative. This work is different from the traditional cross-lingual SA approaches described above in that it makes an assumption about the target language as well. The co-training-based approach allows use of unlabeled target language documents for learning a sentiment predictor for target language (in this case, Chinese). In order to achieve cross-lingual projections, machine translation is used.

5.1.1 Approach

The co-training algorithm is a bootstrapping-based algorithm that begins with a limited set of labeled data and incrementally adds to the labeled data. The authors note that is algorithm has been applied to reference resolution, statistical parsing, etc. The co-training algorithm results in two classifiers: one using source language features (*i.e.* English features) and another using target anguage features (*i.e.* Chinese features). The sentiment prediction for a document is the normalized prediction made by by both the classifiers individually.
The co-training algorithm receives the following corpora as input:

- English labeled documents: Using a MT system, they are translated to Chinese resulting in Chinese labeled documents.
- Chinese unlabeled documents: Using a MT system, they are translated to English resulting in English unlabeled documents.

The English labeled and unlabeled documents are collectively referred to as the *English view* while the Chinese labeled and unlabeled documents are collectively referred to as the *Chinese view*.

The algorithm iteratively expands the labeled documents of both languages and resultantly trains two classifiers, one for each view. The algorithm proceeds as follows:

- A classifier is trained on English sentiment-labeled documents. The English unlabeled documents are given as input to this classifier. Let this classifier be called classifier E.
- A classifier is trained on Chinese sentiment-labeled documents. The Chinese unlabeled documents are given as input to this classifier. Let this classifier be called classifier C.
- The algorithm then selects p positive and n negative most confidently predicted reviews, each from classifier E and classifier C. The English reviews with their obtained labels are added to the English sentiment-labeled corpus; the Chinese review with their obtained labels are added to the Chinese sentiment-labeled corpus.
- In addition to this, the Chinese documents annotated as a result of step 1 are translated to English and added to the English labeled document set as well. The same is done for English documents - they are translated to Chinese and added to the Chinese labeled set.
- The documents just added to the labeled set are removed from the test corpus of both the classifiers.
- The above process is repeated for a pre-determined set of iterations

5.1.2 Experiment setup

The co-training algorithm has the following tunable parameters:

- i: Number of iterations
- p, n: Number of positive and negative reviews that are selected per iteration

A SVM is trained for both the languages using bigram and unigram features. They are represented using term frequencies. For machine translation, Google translate is used. The datasets used for the experiments are as follows:

- The test set consists of 886 Chinese product reviews.
- The English labeled set consists of 8000 product reviews downloaded from Amazon and annotated based on the user ratings provided in the review.
- Chinese unlabeled set used for cotraining consists of 1000 product reviews downloaded from IT168 website.

The co-training algorithm is run for $i = 40$ and $p = n = 5$. The baseline experiments that are compared with the co-training-based approach are as follows:

- SVM(CN): A SVM is trained on Chinese documents. English documents are translated to Chinese.
- SVM(EN): A SVM is trained on English documents. Chinese doucments are translated to English.
- SVM(ENCN1): Both English and Chinese features are used to learn this classifier. During training as well as testing, MT in either directions is required.
- SVM (ENCN2): The output from SVM(CN) and SVM(EN) is combined by averaging the prediction values.
- TSVM(CN) and TSVM(EN): The method uses transductive SVM for learning a classifier on Chinese and English features respectively.
- TSVM(ENCN1): Like SVM(ENCN1), this classifier also uses features from both the languages.
- TSVM(ENCN2): Like SVM(ENCN2), this classifier combines the prediction values of TSVM(CN) and TSVM(EN) by averaging them.

Classifiers described in 1 to 4 do not use unlabeled Chinese set at all. The classifiers described in 5 onwards use unlabeled Chinese set in the context of transductive SVM.

5.1.3 Results

The authors report class-wise and overall precision, recall and F-scores for different classifiers. Co-training-baesd approach outperforms all other classifiers under all heads. SVM and TSVM-based classifiers show comparable performance in all cases. Among classifiers of their own category (*i.e.* TSVM and SVM classifiers), TSVM(ENCN2) and SVM(ENCN2) exhibit best performance.

The authors repeat co-training experiments for varying values of iterations. It is observed that after i=20 iterations, co-training-based approach outperforms transductive SVM. The authors note that $i = 40 - 70$ is a good range of value for umber of iterations.

To understand the influence of p and n on the accuracy of the clasifier generated, the authors vary these parameters as well. It is observed that imbalanced incremental growth (*i.e.* where number of positive and negative documents added at each iteration is not the same) leads to degradation in performance. The authors report that for p = 1, n = 5 and p = 5, n = 1, the accuracy goes on decreasing as the number of iterations increases. On the other hand, for equal number of p and n added at each ieration, the accuracy remains more or less constant beyond 20-30 iterations.

5.1.4 Summary

The paper presents co-training-based approach to perform sentiment classification of Chinese documents. The salient features of this paper are:

- A set of unlabeled documents in the target language are available. The co-training approach makes use of them to iteratively learn classifiers.
- The baseline experiments involve transductive SVM since it also makes use of unlabeled documents that may be used in the test corpus.
- The work reports changing performance of cross-lingual sentiment classification for varying parameters: number of iterations, number of incremental additions per iteration.

Chapter 6

Conclusion & Future Work

This report describes approaches to cross-lingual SA. We first introduced sentiment analysis and cross-lingual SA. With respect to rule-based approaches for cross-lingual SA, we described (a) how a classifier may be adapted for a target language, (b) how existing sentiment lexicons in other languages may be used to create sentiment lexicons in target language and (c) how lexicons may be created using games on crowd-sourcing platforms. For supervised SA, we described how machine translation may be used for cross-lingual projection. Then, we examined how quality of MT affects cross-lingual SA and show that perfect MT may not lead to perfect cross-lingual SA. Additionally, we presented a study that compares cross-lingual SA using MT with in-language SA and show that in-language SA always proves to be a skyline. Alternate approaches that project features across languages to a common feature space have been reported. This could be done using word senses as features or using structural correspondence learning that finds projections using pivot features. Other alternatives are to learn a joint bilingual classifier that predicts common sub-features across the two languages using labeled target language data, or to use a cross-lingual mixture model that uses a generative framework based on probability of word projections. In addition, we also described a co-training based approach that learns classifiers for the source and target language by iterativeely improving them.

Following future directions may be adopted. A comparative study of how alternate MT implementations affect performance of cross-lingual SA may be studied. However, as described in Duh et al. [2011], perfect MT may still lead to unknown features in the training feature space. While sense-based representation mitigates this problem to an extent, alternate approaches to do so may improve the performance of cross-lingual SA. Also, parallel corpora is often used for cross-lingual SA. It may be useful to see how additional low-level alignments can be leveraged for identifying cross-lingual projections as

pointed out by Meng et al. [2012]. Finally, while words may not possess the same polarity across languages and that idioms need to be projected differently, sentences retain their polarity across languages, it is necessary to use novel adaptation approaches.

Bibliography

AR Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. Harnessing wordnet senses for supervised sentiment classification. In *EMNLP*, pages 1081–1091. ACL, 2011. ISBN 978-1-937284-11-4. URL `http://www.aclweb.org/anthology/D11-1100`.

AR Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. Cross-lingual sentiment analysis for indian languages using linked wordnets. In Martin Kay and Christian Boitet, editors, *COLING (Posters)*, pages 73–82. Indian Institute of Technology Bombay, 2012. URL `http://aclweb.org/anthology/C/C12/`.

AR Balamurali, Mitesh M. Khapra, and Pushpak Bhattacharyya. Lost in translation: Viability of machine translation for cross language sentiment analysis. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 7817 of *Lecture Notes in Computer Science*, pages 38–49. Springer, 2013. ISBN 978-3-642-37255-1. URL `http://dx.doi.org/10.1007/978-3-642-37256-8`.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *EMNLP*, pages 127–135. ACL, 2008. URL `http://www.aclweb.org/anthology/D08-1014`.

Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/R09-1010`.

Amitava Das and Sivaji Bandyopadhyay. Towards the global sentiwordnet. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, editors, *PACLIC*, pages 799–808. Institute for Digital

Enhancement of Cognitive Development, Waseda University, 2010. ISBN 978-4-905166-00-9. URL http://www.aclweb.org/anthology/Y10-1092.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *ACL (Short Papers)*, pages 429–433. The Association for Computer Linguistics, 2011. ISBN 978-1-932432-88-6. URL http://www.aclweb.org/anthology/P11-2075.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, pages 417–422, 2006.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. Joint bilingual sentiment classification with unlabeled parallel corpora. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 320–330. The Association for Computer Linguistics, 2011. ISBN 978-1-932432-87-9. URL http://www.aclweb.org/anthology/P11-1033.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. Cross-lingual mixture model for sentiment classification. In *ACL (1)*, pages 572–581. The Association for Computer Linguistics, 2012. ISBN 978-1-937284-24-4. URL http://www.aclweb.org/anthology/P12-1060.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In John A. Carroll, Antal van den Bosch, and Annie Zaenen, editors, *ACL*. The Association for Computational Linguistics, 2007. URL http://aclweb.org/anthology-new/P/P07/P07-1123.pdf.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *CoRR*, cs.CL/0409058, 2004. URL http://arxiv.org/abs/cs.CL/0409058.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007. URL http://dx.doi.org/10.1561/1500000011.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 3077–3081. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/1081.html.

Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL*, pages 1118–1127. The Association for Computer Linguistics, 2010. ISBN 978-1-932432-66-4; 978-1-932432-67-1. URL http://www.aclweb.org/anthology/P10-1114.

Ellen M. Riloff, Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, and Siddharth Patwardhan. Opinionfinder: a system for subjectivity analysis, 2005. URL http://content.lib.utah.edu/u?/ir-main,47569.

Arno Scharl, Marta Sabou, Stefan Gindl, Walter Rafelsberger, and Albert Weichselbraun. Leveraging the wisdom of the crowds for the acquisition of multilingual language resources. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 379–383. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/210.html.

Xiaojun Wan. Co-training for cross-lingual sentiment classification. In Keh-Yih Su, Jian Su, and Janyce Wiebe, editors, *ACL/AFNLP*, pages 235–243. The Association for Computer Linguistics, 2009. ISBN 978-1-932432-45-9; 978-1-932432-46-6. URL http://www.aclweb.org/anthology/P09-1027.

Janyce Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.