Cross-lingual sentiment analysis with machine translation utility of training corpora and sentiment lexica

Demirtas, E.

Award date:
2013

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

Cross-Lingual Sentiment Analysis with Machine Translation

Utility of training corpora and sentiment lexica

Erkin Demirtas

Master Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

Supervisor:

dr. Mykola Pechenizkiy

Committee Members:

dr. Mykola Pechenizkiy
Prof. dr. Paul De Bra
dr. Alexander Serebrenik

Eindhoven, the Netherlands
October, 2013

Abstract

Recent advancements in machine translation foster an interest of its use in sentiment analysis. This thesis investigates prospects and limitations of using machine translation in cross-lingual sentiment analysis.

To perform a sentiment analysis we need to learn linguistic features by either using tools such as part-of-speech taggers, parsers, or basic resources such as annotated corpora or sentiment lexica. We are motivated to study the translation of existing resources in English simply because building such tools and resources for each language requires considerable human effort. This severely limits the implementation of language specific sentiment analysis techniques similar to those developed for English.

Labeled corpora and sentiment lexica are two main resources in the application of sentiment analysis. We translate them to a language with limited resources where we opt to focus on improving classification accuracy when (labeled or raw) training instances are available. In some cases, however, we may not have access to any training data. To address this scenario we explore methods to translate sentiment lexica to a target language as we also try to improve machine translation performance by generating additional context.

For all experiments we work on English and Turkish data which consist of movie and product reviews and we perform two-class (positive-negative) classification -polarity detection in which we discard the neutral class. Consequently, we obtain promising results in polarity detection experiments where we use general-purpose classifiers trained on translated corpora while in this point we remark that dissimilarities between two corpora in different languages should be further studied for better integration of resources. We also find quantitative evidences to suggest that lexica translation is more troublesome since the inherit differences of expressing sentiment between two languages make it harder to preserve the sentiment of words/phrases when translating them from one language to another.

Preface

I owe a deep gratitude to my supervisor Mykola Pechenizkiy for valuable advice and friendly guidance. I am thankful that he supported me in my difficult times as well. I am also grateful to Paul De Bra and Alexander Serebrenik that they accept to be in my thesis committee in a short notice.

I owe a lot to my parents, for everything they have done for me, for all their love and care.

Contents

Abstract i

Preface ii

Contents iii

List of Figures iv

List of Tables v

1 Introduction 1
- 1.1 Motivation & Goals . . . 2
- 1.2 Approach . . . 3
- 1.3 Thesis Organization . . . 4

Preliminaries 5

2 Sentiment Analysis 5
- 2.1 Different Levels of Granularity . . . 6
- 2.2 Computational approaches . . . 7

3 Multi-lingual and Cross-lingual Sentiment Analysis 10
- 3.1 Possible Scenarios . . . 10
- 3.2 A Room for Machine Translation . . . 11
- 3.3 An Example Target Language: Turkish . . . 11

4 Datasets and Tools 13

Experiments 15

5 Corpus-based cross-lingual projections 15
- 5.1 Background & Methodology . . . 15
- 5.2 Results & Discussion . . . 15

6 Introducing additional cross-lingual training data **18**
6.1 Expanding training corpus size with unseen machine translated data 19
6.1.1 Background & Methodology 19
6.1.2 Results & Discussion 20
6.2 Co-training with machine translation 23
6.2.1 Background & Methodology 23
6.2.2 Results & Discussion 23

7 Issues in sentiment lexicon translation **26**
7.1 Background 27
7.2 Methodology 27
7.2.1 Bilingual Dictionary 29
7.2.2 Machine Translation 30
7.3 Results & Discussion 33
7.3.1 Coverage of most frequent words 34
7.3.2 Accuracy of sentiment orientation 34
7.3.3 Effectiveness in Polarity Detection 34

8 Conclusions **38**
8.1 Summary of results 38
8.2 Contributions 39
8.3 Future Work 41

Bibliography **42**

List of Figures

1.1 Text Classification Task 1

2.1 Lexicon and ML centric approaches for learning a sentiment classifier 8

6.1 Study of training set expansion with machine machine. 20
6.2 Feature size comparison for the training set expansion experiment. 21
6.3 Generalization accuracies for the training set expansion experiment. 22
6.4 Co-training experiment setup. 24
6.5 Accuracy comparison for the co-training experiment. 24

7.1 Number of sentiment words compiled from SWN 28
7.2 Desnsity graphs for lexicons compiled from SWN 28
7.3 Number of sentiment words in MT lexicons 31
7.4 Context based translation of SentiWordNet 31
7.5 Number of sentiment words in CONTEXT lexicon 32
7.6 Comparison of lexicons in terms of size 33
7.7 Desnsity graphs for translated lexicons from SWN-weighted 35

8.1 Summary of experiment setups 38

List of Tables

3.1 Example of Agglutination in Turkish 12

4.1 Size of the datasets used in the experimental study. 14

5.1 Features employed for representing the dataset 16
5.2 Results obtained for English and Turkish using different feature representation 16

6.1 Naïve Bayes classification performance 22
6.2 Linear SVC classification performance 22
6.3 MaxEnt classification performance 22

7.1 Summary of classification accuracies and comparison with Denecke [10] . . . 29
7.2 Frequent words coverage 34
7.3 Accuracy in sentiment orientation 34
7.4 Sentiment classification comparisons (Books) 36
7.5 Sentiment classification comparisons (DVD) 36
7.6 Sentiment classification comparisons (Electronics) 37
7.7 Sentiment classification comparisons (Kitchen) 37

Chapter 1

Introduction

Language is one of the most fascinating discoveries of mankind. Whether it is spoken or written, it is our intelligent that reaches an agreement of communication and forms a language to express our ideas, emotions, and all other feelings. As we are in the information age now, people can communicate with each other by means of electronic forms and this results massive amount of resources that re-shape the world of "sentiment analysis". In short, it is basically the study of opinions, sentiments, appraisal, and emotions expressed in text. The term has already been known even before the first computers were invented, but the rapid growth of digital data and widespread information flow stimulate the development of computational methods, as the old-school techniques are not feasible anymore. These developments set aflame with the desire for exploring new challenging problems, never studied before. As attractive as it seems, however, it is not an easy task. Considering the fact that even human annotators may not agree on a sentiment; finding a concise, reliable and accurate way of sentiment analysis requires attentive work on different linguistic features.

Sentiment analysis is a text classification task (see Figure 1.1) where it maps linguistic features to several rules to assess the sentiment of the text. Let $\{f_1, f_2, ..., f_n\}$ is the set of features that can appear in text. These could be the numbers, words, phrases, or even characters itself such as period, comma etc. Let $\{r^1, r^2, ..., r^m\}$ is the set of rules which assert the way features are used to represent the text. For example, if we take the word f_i as a feature and r^1 calculates the word frequency then $r^1(f_i, t)$ gives us the number of times the word f_i appear in text t. Hence in general a text t is represented by the vector

$$\overrightarrow{t} := ((r^k(f_i, t)) \text{ where } k \in 1, ..., m \text{ and } i \in 1, ..., n$$

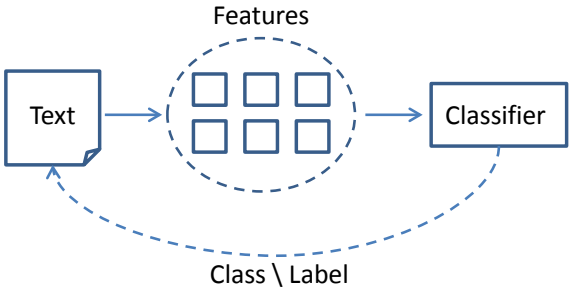


Figure 1.1: Text Classification Task

With this representation of a given text t, a typical sentiment classifier assigns to it the class c* $\in$ {positive, negative, neutral} or {objective, subjective}.

At a high level of abstraction, a natural language processing (NLP) system provides a mapping that specifies how the linguistic structure underlying natural language text, such as parts of speech, or syntactic and semantic relations, is to be uncovered from its surface form. In the early days, this mapping was composed of hand-crafted rules that specified, for example, how words with particular parts of speech fit together in certain syntactic relations. Instead, modern systems for linguistic analysis typically employ highly complex rules that are automatically induced from data by means of statistical machine learning methods. Due to the inherent ambiguity and irregularity of human natural languages, the mapping provided by a high-accuracy linguistic processing system is tremendously complex.

There are several ways in which knowledge can enter a NLP system. At a high level, we identify the following three sources of knowledge:

1. Expert rules: Human experts manually construct rules that define a mapping from input text to linguistic structure. This is typically done in an iterative fashion, in which the mapping is repeatedly evaluated on text data to improve its predictions.

2. Labeled data: Human experts annotate text with the linguistic structure of interest. A mapping from input text to linguistic structure is then induced by supervised machine learning from the resulting labeled data.

3. Unlabeled data: Human experts curate an unlabeled data set consisting of raw text and specifies a statistical model that uncovers structure in this data using unsupervised machine learning. The inferred structure is hoped to correlate with the desired linguistic structure.

We can also use these sources together. The class of combined methods that have received most attention are semi supervised learning methods, which exploit a combination of labeled and unlabeled data to improve prediction.

We discuss more about building a sentiment classifier using such sources in Chapter 2, but before that we explain our focus in sentiment analysis, the motivation behind this thesis, and the research questions we try to address through our experimental study in following sections.

1.1 Motivation & Goals

Construction of NLP tools require large amounts of training data that has been annotated with the linguistic structure of interest, to reach a satisfactory level of performance. The cost of creating these resources manually is so high that such tools are currently lacking for most of the world's languages. Although the volume of non-English sentiment analysis research is increasing, majority of studies in the field still concentrates on English. Many advanced tools developed for English are not available for other languages yet, which strains the applicability of sentiment analysis on these languages.

Cross-lingual or multilingual approaches bring the possibility of building sentiment analysis applications in other languages than English with less effort comparing to the one required for in-language methods. Besides their implications for each language in individual level, these approaches may stand out as a way to resolve bias in news around the world and may contribute to our understanding of global phenomena. Thus, we are motivated to study the translation of existing resources in English by using machine translation capabilities.

In this context, typical resources required to build sentiment classifier are labeled corpus

and sentiment lexicon. The former consists of documents with their annotations of interest i.e. movie reviews marked by their scores out of 10. The latter is basically a compilation of words/phrases which is given a sentiment score or annotation.

In particular, my main **goal** is to investigate whether creating such resources with machine translation is a viable alternative to labor intensive manual annotation tasks. With a series of experiments we seek to find out the advantages or limitations of using machine translation in this manner. On the application side we also use these resources to develop sentiment classifiers in a target language. We can now define our high-level problem definition as follows:

Problem Definition: Given a text in a resource-poor language, determine whether it expresses a sentiment by making projections from a resource-rich language using machine translation.

Following this definition we try to answer the research questions listed below:

- How useful the resources created by machine translation in sentiment analysis? Is it a viable alternative to time-consuming manual annotation task?
- Assuming we have access to additional labeled data from another language, how can we make use of it to improve the classification performance?
- Assuming we have access to additional raw data from another language, how can we make use of it to improve the classification performance?
- Assuming we have no access to any training data in our language of interest, then what are the ways to translate a sentiment lexicon to perform sentiment classification in that language?

1.2 Approach

This study is concerned with the applicability of machine translation as a bridge to overcome language gap while projecting sentiment resources across languages. The approach follows several experiments to realize possible scenarios of building a sentiment analysis system in a resource-poor language using such projections.

Cross-lingual sentiment research mainly focuses on two alternatives. Corpus-based and lexicon-based projections. I replicate some important studies in this field such as building annotated data set with corpus-based projections [4], application of co-training idea to build polarity classifier in target language [40], and generating context for machine translation engine to obtain alternative translations of same word/phrase [25] The results I obtain from these replica studies give me detailed insight to analyze advantages and shortcomings of proposed methods. I also conduct an experiment with different amount of training data taken from language-specific or cross-lingual corpora.

For all experiments, I opt to study the restricted case of review (product or movie) sentiment, where I assume that the document only bears one direction of overall sentiment out of the two-class: positive and negative. Although two-stage approach in classification i.e. identification of subjective text performed before the polarity detection is often beneficial, and finer-grained sentiment definitions could be employed I deliberately keep experiment

setups simple to focus on only the effectiveness of translation approaches and conduct easily reproducible research.

1.3 Thesis Organization

The remainder of this thesis is organized into the following chapters:

Chapter 2 describes the basics of sentiment analysis along with a brief summary of related work.

Chapter 3 introduces possible scenarios for cross-lingual sentiment analysis, argues the use of machine translation, and introduces an example target language - Turkish.

Chapter 4 provides details of data sets and tools used in experiments.

Chapter 5 considers an experiment using corpus-based cross-lingual projections.

Chapter 6 studies cross-lingual polarity detection where (labeled or unlabeled) data in target language is available.

Chapter 7 compares the methods for translation of sentiment lexicons.

Chapter 8 concludes the thesis with a summary of all experiment results, final remarks, contributions and directions for further research.

Some of the material in this thesis in particular the material in chapter 6 were previously published in the following paper:

E. Demirtas and M. Pechenizkiy. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 9:1–9:8, New York, NY, USA, 2013. ACM.

Chapter 2

Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.

There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. There is little agreement on terminology in the literature on subjective language. Common terms used for what we here call sentiment, include opinion, attitude and affect. It is important to note that the terms *opinion mining*, *subjectivity analysis*, *sentiment analysis* have slightly different meanings, but *sentiment analysis* is used in most part of this thesis as a broader term to cover all these definitions for simplicity, and I use the other terms only where their specific meanings are need to be used.

The analysis and processing of subjective language, as manifested for example in sentiments, beliefs and judgments, is a growing area within natural language processing. Although some interest in this area can be traced to the 1960s [37], there has been a surge of interest in the field in the last ten years, primarily thanks to the increasing significance of informal information sources, such as blogs and micro-blogs, user review sites and the booming growth of online social networks; see [30] for a comprehensive overview of this development. Current research suggest that analyzing subjectivity in language is more difficult, compared to more traditional tasks related to content or topicality [20]. Whether this is due to the immature nature of the field or an inherent aspect of the problem has not been settled. However, the inter-annotator agreement is typically quite low for subjective aspects of language, compared to topical aspects, which suggests that subjective language analysis is indeed an intrinsically harder problem.

There is also no agreement on sentiment, subjectivity and opinion definitions but there are variations proposed in literature. [42] defines the term subjective language as referring to aspects of language use related to the expression of private states, such as sentiments, evaluations, emotions or speculations. A private state is characterized by a sentiment, possibly having a polarity of a certain degree, a holder and a topic.

[21] defines the opinion as a quadruple. An opinion is a quadruple, (g, s, h, t), where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed. This definition, although quite concise, may not be easy to use in practice especially in the domain of online reviews of products, services, and brands because the full description of the target can be complex and may not

even appear in the same sentence.

Most research on subjective language has focused on sentiment in isolation, or on sentiment in combination with polarity. Interest has commonly been limited to the identification of sentiment, without any further distinction between different types of sentiments, their topics or holders; and to classification of polarity into the categories of positive, negative and neutral [33]. Thus, even when ignoring the directional aspects of holder and topic, most work has been rather coarse-grained in the characterization of private states. Some notable exceptions are [8, 19, 18] and more recently [17] also study methods for holder and topic identification.

Most work on subjective language analysis has been framed at the document level. Some notable examples are [31, 39] in which polarity classification is applied to movie reviews according to a thumbs up/thumbs down classification scheme. Other examples are [29, 14] who also classify movie reviews, but use a multi-point rating scale instead of a bipolar classification.

2.1 Different Levels of Granularity

Most work in the area disregards the difficult aspect of topic and holder and instead simply analyze a piece of text as being dominated by positive, negative or neutral sentiment. This is the most well-studied scenario in the subjective language analysis literature and is often referred to as sentiment analysis or opinion mining (Pang and Lee, 2008). Typically, this analysis is carried out at the word-level, at the document-level, or at the sentence-level.

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. This level of analysis assumes that each document expresses opinions on a single entity.

Work at this level has been done in [39] on movie reviews. The sentiment polarity of the individual opinion words is computed using a set of seed adjectives whose polarity is previously known and computing the Point-wise Mutual Information score that is obtained between the word to classify and the known word using the number of hits obtained by querying the two words together with the NEAR operator on the AltaVista search engine. The final score obtained for the review is computed as sum of the polarities of the individual opinionated words in the review, from a set of sentences that is filtered according to patterns bases on the presence of adjectives and adverbs.

Another approach at the classifying polarity of sentiment at a document level is presented in [31], where the authors use Naive Bayes machine learning using unigram features and show that the use of unigrams outperforms the use of bigrams and of sentiment-bearing adjectives.

Sentence level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to *subjectivity classification*, which distinguishes sentences that express factual information from ones that express subjective views and opinions. However, one should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions.

Sentiment analysis at the sentence level includes work by [28], where an algorithm based on computing the minimum cut in a graph containing subjective sentences and their similar-

ity scores is employed. [45] uses sentence level sentiment analysis with the aim of separating fact from opinions in a question answering scenario. Other authors use subjectivity analysis to detect sentences from which patterns can be deduced for sentiment analysis, based on a subjectivity lexicon [41, 44].

Aspect level: Aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a *sentiment* and a *target*. The goal of this level of analysis is to discover sentiments on entities and their aspects.

While detecting the general attitude expressed in a review on a movie suffices to take the decision to see it or not, when buying an electronics product, booking a room in a hotel or traveling to a certain destination, users weigh different arguments in favor or against, depending on the "features" they are most interested in (e.g. weight versus screen size, good location versus price). Reviews are usually structured around comments on the product characteristics and therefore, the most straightforward task that can be defined in this context is the feature-level analysis of sentiment. The feature-level analysis is also motivated by the fact that on specific e-commerce sites, reviews contain special sections where the so-called "pros" and "cons" of the products are summarized, and where "stars" can be given to value the quality of a characteristic of a product (e.g. on a scale from 1 to 5 "stars").

Recently, authors have shown that performing very fine or very coarse-grained sentiment analysis has drawbacks for the final application, as many times the sentiment is expressed within a context, by comparing or contrasting with it. This is what motivated in [24] to propose an incremental model for sentiment analysis, starting with the analysis of text at a very fine-grained level and adding up granularity to the analysis (the inclusion of more context) up to the level of different consecutive sentences. The authors showed that this approached highly improved the sentiment analysis performance. The same observation was done in [1] for the task of feature-based opinion mining.

2.2 Computational approaches

Most research on sentiment analysis can be categorized into one of two categories: lexicon-centric or machine-learning centric (See Figure 2.1). In the former, large lists of phrases are constructed, manually or automatically, which indicate the polarity of each phrase in the list. This is typically done by exploiting common patterns in language [16, 32], lexical resources such as Word-Net or thesauri [42, 27], or via distributional similarity [39].

In the machine-learning centric approach, one instead builds statistical text classification models based on labeled data, often obtained via consumer reviews that have been tagged with an associated star-rating [31, 5]. Both approaches have their strengths and weaknesses. Systems that rely on lexica can analyze text at all levels, including the clausal and phrasal level, which is fundamental to building user-facing technologies such as faceted opinion search and summarization [23]. However, lexica are typically deployed independent of the context in which mentions occur, which makes them brittle in the face of domain shifts and complex syntactic constructions [43]. The machine-learning approach, on the other hand, can be trained on the millions of labeled consumer reviews that exist on review aggregation websites, often covering multiple domains of interest [31, 28, 5] The downside is that the supervised learning signal is often at a coarse level, most commonly the document level.

Attempts have been made to bridge this gap. The most common approach is to obtain a

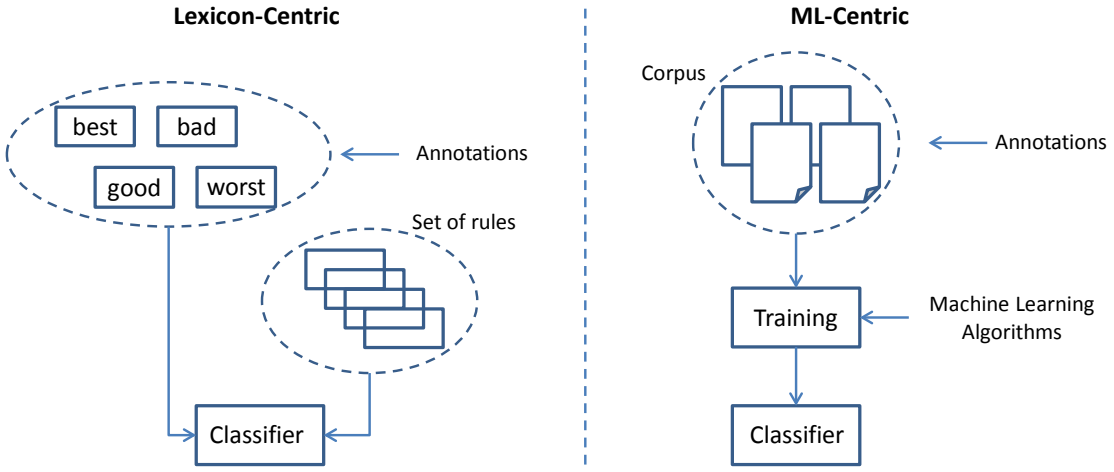


Figure 2.1: Lexicon and ML centric approaches for learning a sentiment classifier

labeled corpus at the granularity of interest in order to train classifiers that take into account the analysis returned by a lexicon and its context [43]. This approach combines the best of both worlds: knowledge from broad-coverage lexical resources in concert with highly tuned machine-learning classifiers that take into account context. The primary downside of such models is that they are often trained on small data sets, since fine-grained sentiment annotations rarely exist naturally and instead require significant annotation effort per domain [41].

We do not provide a summary of statistical machine learning methods in this thesis since it is out of focus. We omit the details of these methods to focus on the part we actually investigate. We use them when necessary as black-boxes to have benchmark results. However, below we put a summary of sentiment lexicons since in Chapter 7 we study to build sentiment lexicons for a resource-poor language.

Sentiment Lexicons The most important indicators of sentiments are sentiment words which are commonly used to express positive or negative sentiments. A list of such words and phrases is called a sentiment lexicon. However they are only part of the story, several other issues have to be addressed as well. A positive or negative sentiment word may have opposite orientations in different application domains. Moreover, a sentence containing sentiment words may not express any sentiment. Sarcasm in sentences also brings additional complexity, and sometimes sentences without sentiment words can imply opinions.

Researchers have proposed many approaches to compile sentiment words. Three main approaches are: manual approach, dictionary-based approach, and corpus-based approach. Since manual approach is labor intensive and time consuming, research is more focused on automated methods.

Dictionary-based Approach: Since most dictionaries list synonyms and antonyms for each word, simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. The advantage of using a dictionary-based approach is that one can easily and quickly find a large number of sentiment words with their orientations. On the other hand sentiment orientations of words collected this way are general or domain and context independent, and as mentioned above many sentiment words have context dependent orientations.

Corpus-based Approach: The corpus-based approach has been applied when a seed list of known sentiment words are in hand, and we are looking for other sentiment words and their orientations from a domain corpus, and also when adapting a general-purpose sentiment lexicon to a new one using a domain corpus for sentiment analysis applications in the domain. However, even in the same domain many words can have different orientations e.g. *"The battery life is long"* (positive) and *"It takes a long time to focus"* (negative). Thus, having domain-dependent sentiment words is insufficient and one still needs to utilize NLP tools to obtain better results.

Chapter 3

Multi-lingual and Cross-lingual Sentiment Analysis

Access to core natural language processing tools is still lacking for most languages, due to the reliance on fully supervised learning methods, which require large quantities of manually annotated training data.

While annotated resources for syntactic parsing and several other tasks are available in a number of languages, we cannot expect to have access to fully annotated resources for all tasks in all languages any time soon. Hence, we need to explore alternatives to methods that rely on full supervision in each target language.

The rationale for cross-lingual learning is that, rather than starting from scratch when creating a linguistic processing system for a resource-poor target language, we should take advantage of any corresponding annotation that is available in one or more resource-rich languages. Typically, this is achieved either by projecting annotations, or by transferring models, from source language to target language.

3.1 Possible Scenarios

Previously there were some authors who tried different approaches to transfer the knowledge in sentiment analysis from English to other languages.

[26] propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon [43] and use two bilingual English-Romanian dictionaries to translate the words in the lexicon. Since word ambiguity can appear (Opinion Finder does not mark word senses), they filter as correct translations only the most frequent words. The problem of translating multi-word expressions is solved by translating word-by-word and filtering those translations that occur at least three times on the Web.

Another approach in obtaining subjectivity lexicons for other languages than English was explored by [4]. To this aim, the authors perform three different experiments, obtaining promising results. In the first one, they automatically translate the annotations of the MPQA corpus and thus obtain subjectivity annotated sentences in Romanian. In the second approach, they use the automatically translated entries in the Opinion Finder lexicon to annotate a set of sentences in Romanian. In the last experiment, they reverse the direction of translation and verify the assumption that subjective language can be translated and thus new subjectivity lexicons can be obtained for languages with no such resources.

[6] experiment with translation from the source language (English) to the target language (Spanish) and then used a lexicon-based approach or machine learning for target language document sentiment classification.

[36] create sentiment dictionaries in other languages using a method called "triangulation". They translate the data, in parallel, from English and Spanish to other languages and obtain dictionaries from the intersection of these two translations.

3.2 A Room for Machine Translation

Broadly speaking we want to cover more languages in sentiment analysis, but given that multilingual analysis is an expensive task, how could we make it possible? Machine translation may reduce the cost by providing access to complex tools that are already exists for English, but there only few works concentrated on this method. In order to validate whether it is a reliable alternative to the traditional methods, we have to see more experiment which considers more languages and focuses on data from different domains. Following this argument we will especially focus on sentiment analysis woth machine translation in Turkish.

[11] presented their opinions about the research of multilingual sentiment classification, and they claimed that domain mismatch was not caused by machine translation (MT) errors, and accuracy degradation would occur even with perfect MT.

Finally, [2] employ fully-formed machine translation systems, also study the influence of the difference in translation performance has on the sentiment classification performance. They report even in the worst cases, when the quality of the translated data is not very high, the drop in performance is of maximum 12%.

Attempts to use machine translation in different natural language processing tasks have not been widely used due to poor quality of translated texts, but recent advances in Machine Translation have motivated such attempts. In Information Retrieval, [34] proposed a comparison between Web searches using monolingual and translated queries. On average, the results show a drop in performance when translated queries are used, but it is quite limited, around 15%.

3.3 An Example Target Language: Turkish

Turkish is very agglutinative language which means most words are formed by joining morphemes together. One word can have many affixes and these can also be used to create new words, such as creating a verb from a noun, or a noun from a verbal root. In Table 3.1 several words that are produced with agglutination on root form of a Turkish verb "gel" (which translates to "come" in English) are presented. It takes different suffixes in each row and the meaning of the word changes dramatically, even it may become a sentence by itself. As seen in Table 3.1, sentiment may be hidden inside the word, and if the current state of art in machine translation is successful enough to differentiate meanings of those words from each other we may as well use this machinery to extract the correct sentiment out of it. Unfortunately, this is very optimistic. In fact when we look at the machine translation results in Table 3.1 we see there are some words which remain same, showing no luck for any translation effort. Although they are rather contrived examples, it proves that data losses happen in the process of machine translation. Luckily, apart from three examples we have translations of

Table 3.1: Example of Agglutination in Turkish

Turkish	English(Google Translate)	English(Actual Meaning)
gel-	come	(to) come
gelebil-	can come	(to) be able to come
gelme-	coming	not (to) come
geleme-	geleme	(to) be unable to come
gelememi	gelememi	Apparently (s)he couldn't come
gelebilecek	can	(s)he'll be able to come
gelmeyebilir	may not be	(s)he may (possibly) not come
gelebilirsen	can come	if you can come
gelinir	is reached	(passive) one comes, people come
gelebilmeliydin	gelebilmeliydin	you should have been able to come
gelebilseydin	could come	if you could have come
gelmeliydin	You should have	you should have come

other words. They might not be perfect translations, but most of them preserved the sentiment by adding corresponding modal verbs to the translation. The stress on the probability of the event in the word "gelmeyebilir", for example, is preserved in its translation "may not be" although the verb itself is disappeared strangely. Following these examples we can argue that machine translation performance of the system and the sentiment accuracy is not linearly correlated i.e. they are not necessarily improved with the same pace when translation accuracy increases. Furthermore one can expect more reliable results from its application in sentiment analysis than its accuracy in translation. If so, utilization of machine translation for sentiment analysis may well be worth a look.

Chapter 4

Datasets and Tools

Datasets:

The following datasets are used in the experiments:

English movie reviews[1]: We use the sentence polarity data which was first introduced by [29]. This data consists of 5331 positive and 5331 negative snippets each containing roughly one single sentence. Reviews are gathered from Rotten Tomatoes web pages for movies released in 2002. They classify reviews marked with *fresh* are positive, and those marked with *rotten* are negative.

English multi domain product reviews[2]: This dataset was first introduced by [5]. It contains product reviews taken from Amazon.com from many product types. For our experiment we use a benchmark dataset which they constructed from four categories (books, dvd, electronics, and kitchen appliances) each consisting of 1000 positive and 1000 negative reviews.

Turkish movie reviews: We collect Turkish movie review dataset from a publicly available website[3]. In order to reach same size with the English dataset we restrict this dataset with 5331 positive and 5331 negative sentences. In this website, reviews are marked in scale from 0 to 5 by the same users who made the reviews. We consider a review positive if its rating is equal to or above 4, and negative if it is below or equal to 2.

Turkish multi domain product reviews: After building Turkish movie reviews dataset, we also collect Turkish product reviews from a Turkish e-commerce website[4] to conduct our training set expansion experiment with reviews from different domains. We constructed another benchmark dataset also consisting reviews from books, DVD, electronics, and kitchen appliances categories to use them along with English product reviews. In this website, reviews are marked in scale from 1 to 5, and majority class of reviews converges to 5, that's why we have to consider a small amount of reviews marked with 3 stars as bearing a negative sentiment to be able to construct a balanced set of positive and negative reviews. It has 700 positive and 700 negative reviews for each of the four categories in which average rating of negative reviews is 2.27 and of positive reviews is 4.5.

Turkish movie reviews and multi domain product reviews datasets are available at `http://www.win.tue.nl/~mpechen/projects/smm/#Datasets`

[1] The dataset is available at `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

[2] The dataset is available at `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

[3] `http://www.beyazperde.com`

[4] `http://www.hepsiburada.com`

Table 4.1: Size of the datasets used in the experimental study.

	English Movie Reviews	Turkish Movie Reviews	English Product Reviews				Turkish Product Reviews			
			Books	DVD	Electronics	Kitchen Appliances	Books	DVD	Electronics	Kitchen Appliances
Positive	5331	5331	1000	1000	1000	1000	700	700	700	700
Negative	5331	5331	1000	1000	1000	1000	700	700	700	700

Tools:

For the most part of experiments we use python with NLTK[5] and scikit[6] libraries. When we build general-purpose classifiers we use scikit implementations of Naive Bayes, SVM, and Maximum Entropy algorithms.

In Chapter 5, we use the tool -LightSIDE[7] to extract feature space and later to train classifiers.

We also build a crawler to construct the Turkish datasets. It first looks for all the movies or products, and lists them. Then it extracts any review written for each product while preserving the score given to those reviews by their authors. For experiment setup we do additional filtering to be able obtain balanced corpora which have equal amount of positive and negative instances.

In order to translate existing resources in English to Turkish, we use two translation engines interchangeably: Google Translate and Yandex Translator.

We use Microsoft N-gram Services [8], an online N-gram corpus built from Web documents in order to generate most frequent bigrams in Chapter 7.

Also in Chapter 7, we use a Turkish NLP library -Zemberek[9] to filter out the English words that could not be translated to Turkish by machine translation engines.

[5] http://nltk.org/
[6] http://scikit-learn.org/
[7] http://www.cs.cmu.edu/~emayfiel/side.html
[8] http://web-ngram.research.microsoft.com/info/
[9] http://zemberek.googlecode.com/svn/trunk/

Chapter 5

Corpus-based cross-lingual projections

5.1 Background & Methodology

This experiment follows similar methods to those presented by [2]. They employed fully formed machine translation systems to translate data from English to three languages - French, German and Spanish. They also created a gold standard for all there languages to measure translation quality. Different than their methodology, I used 5-fold cross-validation for evaluation. Datasets used for the analysis are also different as they use the data provided for English in the NTCIR 8 Multilingual Opinion Analysis Task, I used the sentence polarity data which was first introduced by Pang and Lee in 2005. This data consists of 5331 positive and 5331 negative snippets each containing roughly one single sentence. As a machine translation system Google Translate was employed to translate this dataset into Turkish. Then labels from original data was matched with Turkish translation to finally construct the manually annotated Turkish corpus. After completing this step, I built classifiers using original data set in English to provide the baseline for machine translation performance, and I repeated same procedure by using machine translation of the same data set. While extracting feature space and later training the classifiers I used the LightSIDE[1].

5.2 Results & Discussion

I obtained 5579 unigram, and 5975 bigram features from the original set. Later I translated the corpus into Turkish and extracted the labels from the original data so as to construct a manually annotated Turkish corpus which has 5616 unigram, and 3614 bigram features (See Table 5.1). Since machine translation mostly depend on word by word translations and not so good on phrase level, number of bigram features in Turkish data ended up being relatively small.

After completing feature extraction, it now comes to the choice of algorithms to train the machine learning models. For this purpose I employed two different algorithms: SMO implementation of Support Vector Machine, and Naive Bayes. Since Balahur et al. have also presented results for SMO, one can compare its relative performance in Turkish comparing to

[1] http://www.cs.cmu.edu/~emayfiel/side.html

Table 5.1: Features employed for representing the dataset

Language	Nr. of unigrams	Nr. of bigrams
English	5579	5975
Turkish	5616	3614

French, German, Spanish. Since the sentence polarity dataset includes only opinions about different movies, it has relatively small amount of noise. Hence, I also choose to employ Naive Bayes method to measure classification accuracy since it performs better when noise level is low. In fact my assumption about the dataset seems to be valid as Naive Bayes outperforms SMO in classification accuracy.

Table 5.2: Results obtained for English and Turkish using different feature representation

Language	Feature Representation	Naive Bayes	SMO
English	Unigram	77.4151	73.7854
English	Unigram+Bigram	77.9122	73.7104
Turkish	Unigram	75.2017	71.2343
Turkish	Unigram+Bigram	75.5768	71.0842

Results is presented in Table 5.2. They are very promising as they shows only 2-3% drop in accuracy which is close to its performance in German. [2] presented up to 3% drop in performance for German, and even in some settings a 1-2% gain comparing to English while in Spanish and French the drop in performance is of maximum 12% in the worst case.

After obtaining statistical results, I take some samples from the data set for further investigation and found out couple of interesting examples which give an idea of the machine translation performance in terms of handling negation in sentences:

Examples:

```
pattern: [language] - "sentence" - predicted - actual

[en] - "although i didn't hate this one , it's not very good either.
it can be safely recommended as a video/dvd babysitter ." - NEG - POS
[tr] - "ben bu bir nefret yoktu, o da ok iyi degil. gvenli bir sekilde
bir video/dvd ocuk bakicisi olarak tavsiye edilebilir." - NEG - POS
```

There some examples like above where classification in source language (English) was wrong, most of the time the problem remains when classifier handles its Turkish translation as I expected. Hence the noise in the actual data, or sentences in mixed sentiment remains the foremost reason for poor accuracy in both language.

```
[en] - "interesting , but not compelling ."  - NEG - NEG
[tr] - "ilgin, ama zorlayici degil." - POS - NEG

[en] - "any film that doesn't even in passing mention political prisoners ,
poverty and the boat loads of people who try to escape the country is less
a documentary and more propaganda by way of a valentine sealed with a kiss ." -NEG-NEG
[tr] - "siyasi tutsaklar, yoksulluk ve yurtdisina kamaya alisan insanlar
```

```
tekne ykleri sz geerken bile degil herhangi bir film az bir pckle mhrlenmis
bir sevgiliye yoluyla bir belgesel ve daha propagandadir." -POS-NEG
```

Those two examples above show wrong classification in Turkish even though they were correct in English. It's hard to say by just looking at individual examples whether it happens because of negation effect in original sentences were diminished or its simply because of the difference in polarity value of corresponding words. The former translation is close to perfect due to its simple grammatical structure, but still its classification is wrong. This might lead us to second alternative (difference in polarity value of corresponding words). In order to speak more confidently we have to see translation from Turkish to English and apply a negation detection mechanism while processing sentences, but it seems from the samples that machine translation is doing well to preserve meaning of negation where sentences are not complex and relatively short. One would argue, therefore, machine translation may become a viable alternative to get rid of high cost of constructing complex NLP tools if the domain is more close to the daily language where sentences are relatively short and expressing simple mood. However, here by no means the definition of daily language contains the language used in social networks as they inherits grammatical inconsistencies, and we think that translation of such dialect with full of abbreviations poses more difficulty on machine translation engines.

Chapter 6

Introducing additional cross-lingual training data

In this experiment, we investigate prospects and limitations of machine translation in sentiment analysis for cross-lingual polarity detection task. We focus on improving classification accuracy in a cross-lingual setting where we have available labeled training instances about particular domain in different languages. We use movie review and product review datasets consisting of polar texts in English and Turkish (see Chapter 4 for more details about the datasets).

The two goals of the experiments are to investigate 1) whether expanding training size with new machine translated instances taken from another corpus improves classification accuracy for the original corpus and 2) whether co-training with machine translation addresses cross-lingual polarity detection.

In general if a text is classified as being subjective, we determine whether it expresses a positive or negative opinion. Structured information available in on-line movie reviews helps us in this regard to eliminate neutrality class as we can rely on user's rating associated on his/her review. We can detect polarity of a subjective review, therefore, based on classified instances on beforehand. However, in the real operational settings we would need to have a subjectivity detection mechanism or three-class polarity detection problem formulation for handling neutral messages. To keep the focus we experiment only with polar messages being either positive or negative.

We can consider cross-lingual sentiment classification as a special case of cross-domain classification settings since even two sources from different languages are from same domain they naturally represent different perspective with respect to cultural biases, hidden sentiments etc. We are tempted to explore how much these differences affect classification performance in a set of movie reviews as it may give hints about applicability of cross-domain classification research on cross-lingual sentiment analysis. We also want to see empirical evidences of introduced machine translation noise in sentiment classification and how much it puts a pressure on potential benefits of having a bigger training set which is expanded with machine translated instances.

We consider two distinct machine translation application scenarios. In the first scenario we simply use machine translation to use labeled instances in Turkish for expanding the training set in English considered as the target language for polarity detection.

In the second scenario we consider the co-training approach as viable alternative to lever-

age machine translated data as it was proposed in [40]. Although we construct labeled Turkish movie and product reviews during our research, for the co-training approach we regard those reviews as unlabeled to be able to setup the similar experimental settings (yet allowing for expanding the evaluation scenarios) and compare our findings with results reported in [40].

In this experiment we use four datasets that we introduced in Chapter 4: English movie reviews, English multi domain product reviews, Turkish movie reviews, and Turkish multi domain product reviews.

6.1 Expanding training corpus size with unseen machine translated data

6.1.1 Background & Methodology

[3] report an improvement in classification accuracy when using out-of-language features, yet our work differs from that in couple of major aspects. Our focus is polarity detection, rather than subjectivity analysis which they investigate. Moreover, their training set is only based on the machine translation of an English corpus, and they do not study how to make use of a new dataset from another language in training set.

A number of approaches have been proposed for polarity detection, including Prior Polarity classification (also with use of an opinion lexicon such as SentiWordNet[1], WordNet-Affect[2] or SenticNet[3]), statistical methods such as support vector machines, neural networks, and Naive Bayes among others. Aspect-based methods are introduced to spot more accurate sentiments on entities and their aspects. New approaches relying on semantic relationships in natural language concepts are also investigated under the concept-level sentiment analysis [7]. In our study we use three popular general purpose classification techniques; Naive Bayes, Support Vector Machines (Linear SVC), and Maximum Entropy (MaxEnt) classification.

As we have labeled datasets in English and Turkish, we can immediately apply any of the supervised learning approaches to build monolingual sentiment classifiers for both languages. At this point, however, we can also investigate a way of improving classification accuracy of a monolingual classifier for the target language using annotated sources in different languages together. Previously a special case of this question was studied in [3], i.e. a pseudo parallel corpora constructed by machine translation services was used, and the focus was on subjectivity analysis. Their study suggested that the subjectivity classification accuracy can be increased by using features drawn from multiple languages. Our first experiment setting follows the idea of using multiple corpora in different languages but in a more generic way as we do not restrict these corpora to be parallel.

For this experiment, we prepare three types of training sets named as *control*, *machine translated*, and *Turkish machine translated* sets. The control set consists of only reviews from the English dataset. In order to measure the effect of machine translation (quality) we construct machine translated set which consists of reviews from English dataset as well, but then they first translated to Turkish and again back to English just to add artificial translation noise to their original form. Finally, we prepare Turkish machine translated set by compiling reviews from Turkish dataset which are translated to English. For all machine

[1] http://sentiwordnet.isti.cnr.it/
[2] http://wndomains.fbk.eu/wnaffect.html
[3] http://sentic.net/downloads/

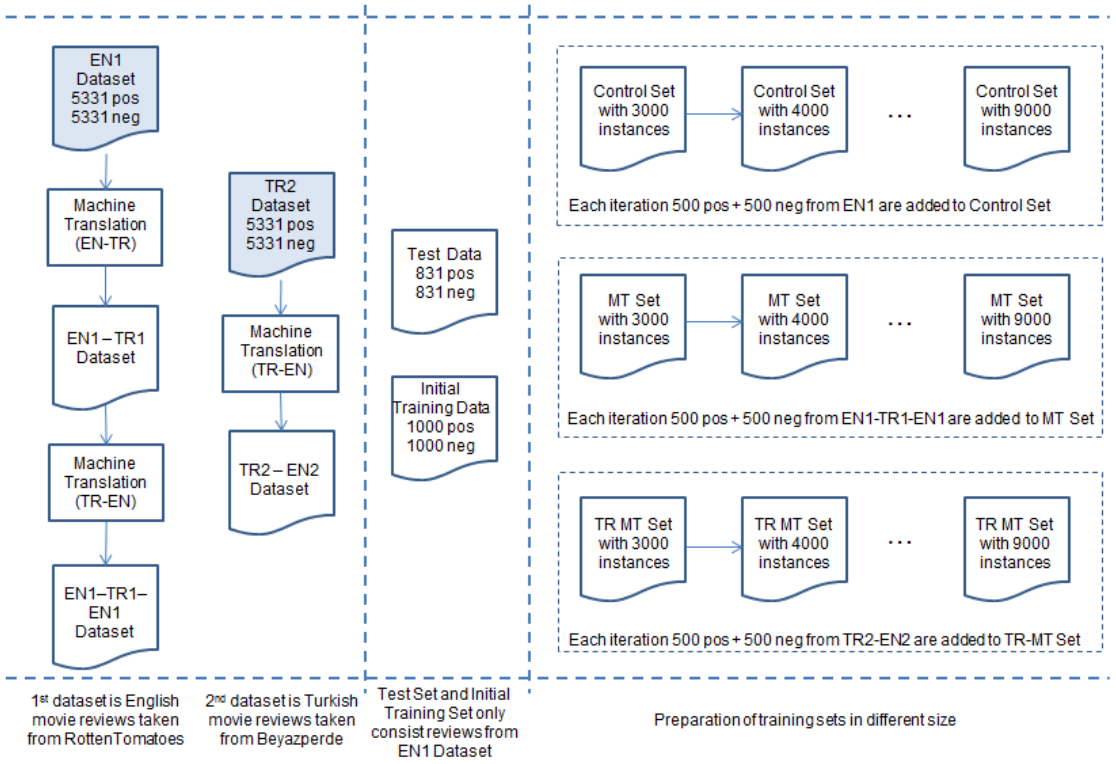


Figure 6.1: Study of training set expansion with machine machine.

translation processes we use Google Translate service.

As Figure 6.1 shows, we first sample 1000 (400)[4] positive and 1000 (400) negative reviews from English movie reviews dataset to run the first iteration of the experiment for both training sets. Then, in every next iteration we increase the size of three training sets by adding 500 (100) positive and 500 (100) negative reviews taken from their respective sources. The test set is constructed from 831 (200) positive and 831 (200) negative English reviews that are never used in the training phase.

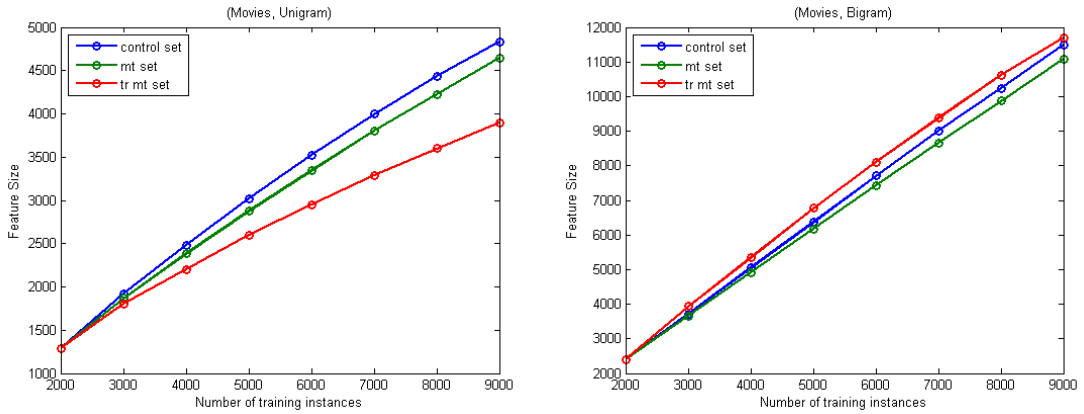
6.1.2 Results & Discussion

For the training set expansion experiment we present our results in terms of two metrics. First, we measure the feature size increase as we keep adding new instances to the training sets.

The two graphs in Figure 6.2 show feature size change of movie reviews datasets in which our training sets are represented by unigram and unigram plus bigram features respectively. We observe an interesting behavior of the feature size change in Turkish machine translated set. Despite its slope is smaller in case of unigram feature representation, when we look at bigram representations it produces more features than the any other set does. Relative poor increase in unigram feature size can be explained by the data loss happened during machine translation as such a number of Turkish words could not be translated to English. On the

[4]numbers in parentheses refer to the setting for product review datasets; without parentheses - to the moview review dataset

other hand, machine translation introduces some amount of noise as well which portrays itself by producing a vast number of meaningless bigrams.



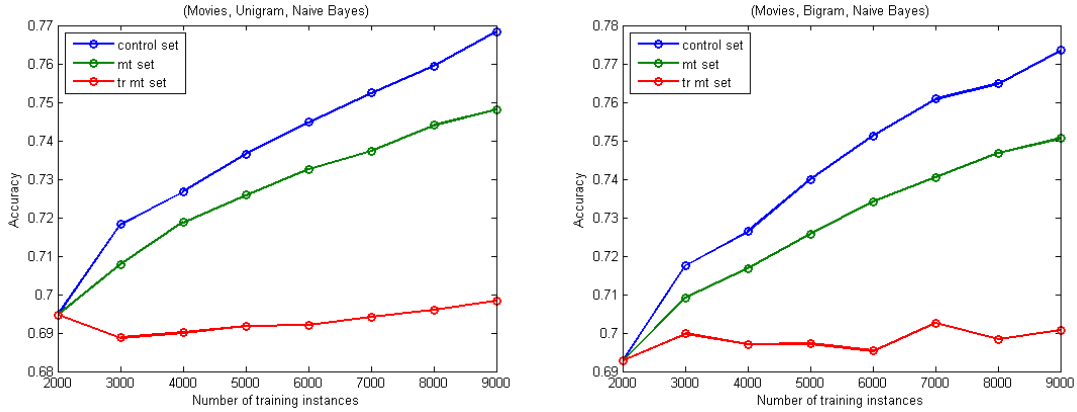
(a) Unigram

(b) Unigram + Bigram

Figure 6.2: Feature size comparison for the training set expansion experiment.

Accuracy results of the Naive Bayes classifiers on movies reviews datasets are summarized in Figure 6.3. We can observe some interesting results. First, consistent with our expectation, expanding training size by adding new instances from the same corpus improves the overall accuracy. This behavior can be noted following the control set results for both graphs in Figure 6.3. Machine translation set slightly under-performs than the control set due to the negative effect of machine translation quality, and this difference tends to increase slightly as we add more machine translated sentences to the training set. Nevertheless, the overall effect of machine translation in this case is positive. We can observe 5% increase in accuracy. The results corresponding to the use of Turkish machine translated set (red line fluctuating between 69% and 70%) clearly shows that naive cross-lingual training set expansion does not improve the generalization performance of polarity detection, although we do gather more features from new instances translated from Turkish movie reviews. This problem refers to cross-domain classification as we can regard new features from Turkish reviews as ones from another domain which is not really immediately helpful to classify the test instances taken from the English dataset. These results suggest that an application of resolving cross-corpora dissimilarity may help to utilize labeled instances taken from another language in cross-lingual sentiment analysis.

This behavior of Naïve Bayes classifier is very similar for Linear SVC and MaxEnt classifiers, and it also generalized to all five datasets we experimented with. The summary of the classification performance is given in Tables 6.1, 6.2 and 6.3. In each table in the first column we give the baseline performance on the initial training data, and the following three columns show the absolute increase (or decrease) in the classification accuracy after the additional training data was added in full according to one of the three setups. We can see from the tables that expanding the training set with additional labeled instances from the same source helps to improve the classification performance and from the different source - does not, and in fact on three datasets even deteriorates the performance.



(a) Unigram

(b) Unigram + Bigram

Figure 6.3: Generalization accuracies for the training set expansion experiment.

Table 6.1: Naïve Bayes classification performance

	Initial accuracy	Control set	MT set	TR MT set
Movies	69.5	+10.6	+7.7	+0.5
Books	72.4	+9.2	+8.6	-0.7
DVD	76.0	+4.6	+1.5	-1.1
Electronics	73.0	+8.1	+9.6	-8.6
Kitchen	75.9	+7.2	+8.7	-6.3

Table 6.2: Linear SVC classification performance

	Initial accuracy	Control set	MT set	TR MT set
Movies	66.0	+11.3	+8.2	+0.5
Books	66.6	+11.1	+14.0	+0.3
DVD	70.3	+7.7	+8.0	-2.7
Electronics	72.4	+7.2	+5.0	-8.0
Kitchen	70.0	+12.3	+11.1	-2.7

Table 6.3: MaxEnt classification performance

	Initial accuracy	Control set	MT set	TR MT set
Movies	68.2	+11.0	+8.8	+0.4
Books	68.7	+12.8	+12.4	+1.8
DVD	71.8	+9.5	+9.6	+1.1
Electronics	74.0	+9.5	+8.0	-7.7
Kitchen	72.4	+12.7	+12.2	-2.2

6.2 Co-training with machine translation

6.2.1 Background & Methodology

In this part we investigate the approach for sentiment classification proposed by [40] who constructs a polarity co-training learning system by using the multi-lingual views obtained through the automatic translation of product-reviews into Chinese and English. While [40] provides empirical evidence that leveraging cross-lingual information improves sentiment analysis in Chinese over what could be achieved using monolingual resources alone, it does not provide any results tested on samples taken from English dataset. Thus, as we show in our experimental study, the conclusions from the reported results in [40] should be interpreted with care.

We use movie reviews instead of product reviews, and we experiment with Turkish-English language setting while Wan uses Chinese-English. These are mostly practical changes in the framework, however, we test combined classifier with reviews taken from both Turkish and English datasets whereas Wan only present results based on test data containing Chinese texts only.

As we can see in Figure 6.4, training input is the labeled English reviews and some amounts of unlabeled Turkish reviews. The labeled English reviews are translated into labeled Turkish reviews, and the unlabeled Turkish reviews are translated into unlabeled English reviews, by using Google Translate. Therefore, each review is associated with an English version and a Turkish version. The English features and the Turkish features for each review are considered two independent and redundant views of the review.

The co-training Algorithm 1 is then applied to learn two classifiers.

The English and Turkish terms (features) used in our study include unigrams; the feature weight is simply set to term presence following the bag-of-words model. The output value of the Naive Bayes classifier for a review indicates the confidence level of the review's classification. In the training phase, the co-training algorithm learns two separate classifiers: C_{en} and C_{tr}. Therefore, in the classification phase, we can obtain two prediction values for a test review, and the average of these values is used as the overall prediction value of the review.

6.2.2 Results & Discussion

Co-training experiment results give us insightful details to compare our findings with the ones reported in [40]. In his paper, Wan evaluates the co-training algorithm by classifying labeled Chinese reviews that are taken from same website and which he used in training phase. We present our results based on labeled Turkish movie reviews corresponding to his labeled Chinese reviews, but also the results based on labeled English movie reviews that are discarded from the training phase. Figure 6.5 confirms findings reported in [40]: tested on labeled Chinese product reviews the combined classifier performs the best and overall accuracy for all classifiers increases in each iteration of co-training. However, co-training approach fails to improve classification accuracy tested on samples from English dataset as we run the algorithm for multiple iteration. For all classifiers (Turkish, English, and combined) we get the highest accuracy in the first iteration as they do not get better with more iterations. Since proposed co-training approach leverages only unlabeled Chinese reviews (in our work these are replaced by unlabeled Turkish reviews) it resembles semi-supervised learning that aims to increase the classification performance with the aid of some unlabeled data in a language which is the same as the language of the test set. Therefore most of the performance gain

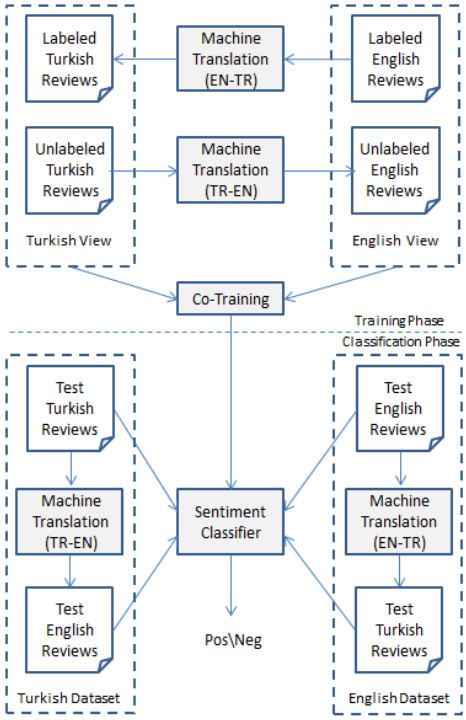
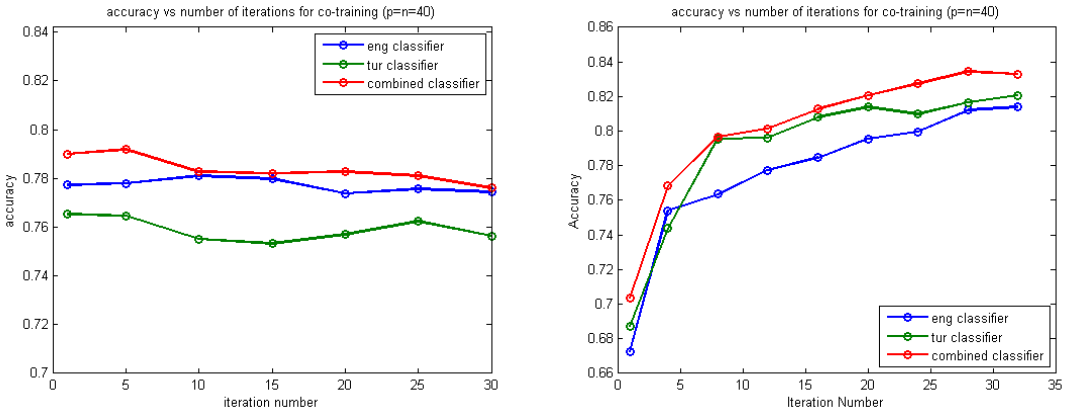


Figure 6.4: Co-training experiment setup.



(a) English dataset

(b) Turkish dataset

Figure 6.5: Accuracy comparison for the co-training experiment.

Algorithm 1 Co-training two classifiers

Input: F_{en} and F_{tr} are redundantly sufficient sets of features, where F_{en} represents the English features, F_{tr} represents the Turkish features, L is a set of labeled training reviews, U is a set of unlabeled reviews
Output: two classifier C_{en} and C_{tr}
for $i \in \{1, 2, \cdots, k\}$ **do**
    Learn the first classifier C_{en} from L based on F_{en}
    Use C_{en} to label reviews from U based on F_{en}
    Choose p positive and n negative the most confidently predicted reviews E_{en} from U
    Learn the second classifier C_{tr} from L based on F_{tr}
    Use C_{tr} to label reviews from U based on F_{tr}
    Choose p positive and n negative the most confidently predicted reviews E_{tr} from U
    Removes reviews $E_{en} \cup F_{tr}$ from U
    Add reviews $E_{en} \cup E_{tr}$ with the corresponding labels to L
end for
return C_{en}, C_{tr}

presented in [40] is likely due to semi-supervised learning rather than the aid of the English classifier.

Chapter 7

Issues in sentiment lexicon translation

Use of sentiment lexicons is one of the common methods in sentiment analysis. It aims to estimate the sentiment expressed in a text by using the polarity (sentiment orientation) of the words. Early works such as [16] relate to this problem first creating a sentiment lexicon in an supervised manner, and then applying a clustering method to determine the polarity of adjectives, where [32] uses bootstrapping in conjunction with an initial high-precision classifier to learn subjective expressions. A recent survey [22] summarizes three main methods for compiling a sentiment lexicon: manual approach, dictionary-based approach, and corpus-based approach. Manual approaches are very costly and time consuming, thus they are often combined with automated methods to build such a lexicon. Dictionary-based approaches work by expanding a small set of seed words with the use of a lexical resource such as the WordNet[1]. The main drawback of these approaches is that the resulting lexicon is not domain specific. Corpus-based approaches can overcome these problems by learning a domain-specific lexicon using a domain corpus of labeled reviews.

Number of sentiment lexicons such as SentiWordNet[2], WordNet Affect[3], SenticNet[4] have already been used in sentiment classification tasks in English. In this work we only experiment with SentiWordNet but one can use other available lexicons as well. [12] built the SentiWordNet, a lexical resource in which each synset of WordNet is associated with a negative polarity, a positive polarity, and an objective polarity to indicate its neutrality.

Contrary to the case in English, sentiment lexicons in other languages are limited or even unavailable because of their high development cost. An alternative to this approach is to transfer an available sentiment lexicon in English to another language in an automated manner. In this work we investigate possible ways to transfer a sentiment lexicon (SentiWordNet) from English to Turkish, and consequently analyze the quality of translated lexicons and their effectiveness in multi-domain sentiment classification.

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

[2] http://sentiwordnet.isti.cnr.it/

[3] http://wndomains.fbk.eu/wnaffect.html

[4] http://sentic.net/downloads/

7.1 Background

Research in building sentiment lexicons for a target language follows mainly two approaches: translation methods or bootstrapping methods. Bootstrapping methods are dependent on semantic resources. They often start with a seed lexicon or network and expand it using available semantic relationships. [15] finds semantic orientation of foreign words based on connection between words in the same language as well as multilingual connections. It assumes the existence of resources (e.g. WordNet, seeds, etc) that often do not exist in foreign languages. [35] follows a graph-based method which looks for similarity in a seed network and tries to induce new relations.

Mihalcea et al. [26] translates an English subjectivity lexicon to Romanian using two bilingual dictionaries. In the same work they compare the performance of resulting lexicon in subjectivity classification with the lexicon they construct using a parallel corpora and they report much lower scores (F-measure 47.93% against 67.85%) when they used the former lexicon. [19] builds rule-based classifiers using a translated lexicon to perform subjectivity analysis in German emails. They use WordNet and machine translation to construct a subjectivity lexicon in German.

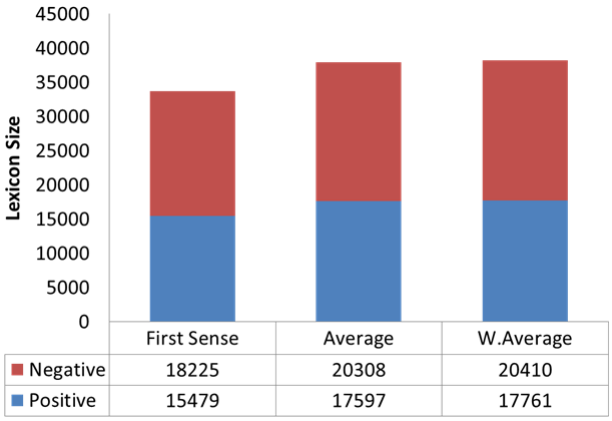
Recently [25] proposes a context based machine translation method to translate a subjectivity lexicon from English to Chinese. Basically they try to put each English sentiment word into a context to generate different phrases which prompts the machine translation engine to return alternative translations for the same word. For this task they use three approaches: looking for frequent collocations, generating coordinated phrases, and placing a punctuation mark at the end of each word. They report an increase in coverage, slight improvement in lexicon precision as well as better results in sentiment classification. However, their work is limited to use of one translation engine, only provides accuracy results in the classification task where they did not experiment with datasets from multiple domains. Moreover they did not report separately how well each proposed approach performs for generating alternative translations.

While we try to address aforementioned shortcomings of [25], we also compare classification results with general purpose classification techniques i.e. Naive Bayes, SVM, and Maximum Entropy. Besides our experiments in Turkish, we also report the accuracy of our rule-based classifier on English reviews and compare them with the benchmark results published in [10].

7.2 Methodology

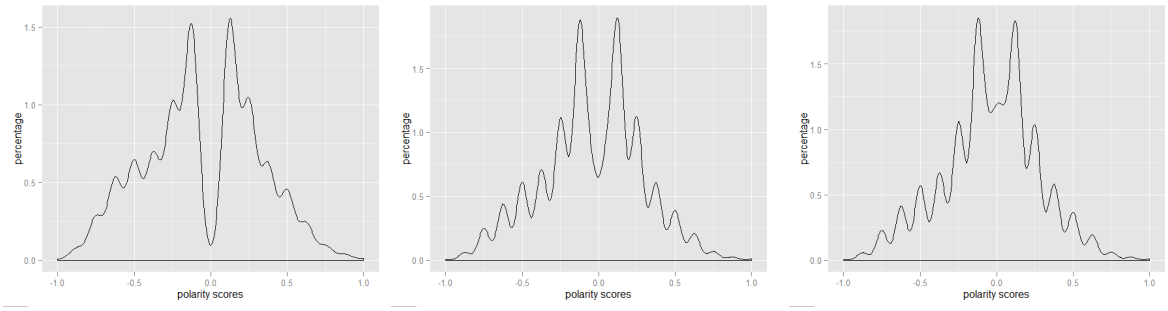
Given that the availability of a sentiment lexicon in English, we build number of sentiment lexicons for the target language: Turkish. We aim to evaluate each translation method in terms of coverage, precision and accuracy metrics. Below we first describe the translation process, then assess the quality of translated lexicons, and finally report their performance in polarity detection.

Bilingual dictionaries and machine translation engines are common tools to translate a word to another language. We use these techniques to compile a sentiment lexicon in Turkish, but besides these benchmarks we also perform context-based machine translation as proposed in [25]. These approaches might not be always correct in translation from source lexicon or some words could not be translated at all, and it refrains us from building a large lexicon



	First Sense	Average	W.Average
Negative	18225	20308	20410
Positive	15479	17597	17761

Figure 7.1: Number of sentiment words compiled from SWN



(a) First Sense (b) S. Average (c) W. Average

Figure 7.2: Desnsity graphs for lexicons compiled from SWN

but even worse, some important aspects of a word's sentiment might be lost in translation because of inherit differences in languages.

In this experiment we try to translate a publicly available sentiment lexicon: SentiWordNet. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. We first determine the final sentiment score of a synset by subtracting negativity score from positive, and thus each synset has a unique sentiment score in between -1 and 1. Since we do not have a tool to resolve word ambiguity in Turkish, we have to compute only one sentiment score for each word as correct as possible. We test three heuristics to determine a final sentiment score from multiple synsets of the same word and size comparison of all three lexicons can be seen in Figure 7.1. First heuristic is to use only the first sense scores for each word as they usually reflect most common usage. Although this might seem reliable assumption, it is not always the case. The word "hot" has 21 synsets and its first adjective sense sentiment score is 0, namely not bearing any sentiment. However, we frequently use the word "hot" both in negative and positive meanings. Moreover using the first sense score for each word is more error prone to the possible mistakes that SWN has. The word "dull" is used for "lacking in liveliness or animation" in its first sense but mistakenly given positivity score of 0.375 and negativity score of 0.25 which ends up a total positive sentiment score (0.125) which is wrong by default because in fact the first meaning of "dull" bears a negative sentiment in most context.

Table 7.1: Summary of classification accuracies and comparison with Denecke [10]

	Books	Dvd	Electronics	Kitchen
Denecke	54	59	65	58
First Sense	59.1	61.4	62.2	64.2
S.Averaged	63.3	65.9	**69.3**	**70.5**
W.Averaged	**65.4**	**66.8**	68.2	69

Second we use the simple average of sentiment scores from each synset of the same word. With this approach we compiled 20308 positive and 17597 negative sentiment words which points 11.08% increase in lexicon size comparing the first-sense approach. Also with this method polarity score of "dull" calculates to -0.059, a more reasonable number comparing to 0.125. Finally we go forward by taking weighted average of sentiment scores, and we assign decreasing weights to synsets: if it is the first meaning then its score gets the maximum weight, and if it is last meaning then its score gets the minimum weight. This method yields 20410 positive and 17761 negative sentiment words which are slightly more than the size of simple averaged lexicon. We also extract density graphs for all three lexicons to visualize the distribution of polarity scores in Figure 7.2. These graphs reveal that taking averages (simple or weighted) mostly brings words with neutral first-sense into the lexicon as they were omitted in the lexicon complied by first-sense approach.

[10] publishes a number of sentiment classification results which uses SWN lexicon in a rule-based sentiment classifier and classifies multi domain product reviews first introduced by [5]. We also built a simple rule-based classifier, and tested on the same dataset. We classify a text by adding up polarity scores of the words appear in it. We use same methodology later in classifying Turkish reviews but for now we aim to see which of the three lexicons (first-sense, s. averaged, w. averaged) performs better in polarity detection and where do they stand comparing the results in [10]. In Table 7.1 we provide a summary of classification results in terms of accuracy. It shows that both SWN-averaged, and SWN-weighted lexicons outperforms the results provided by Denecke.

Since weighted average approach compiles the biggest lexicon among all three, and its polarity detection performance is among the best with the simple average approach, from now on we only use weighted averaged SWN lexicon to build a Turkish sentiment lexicon, and whenever we refer to SWN-weighted, we mean the lexicon compiled by weighted average approach. Having determined the benchmark lexicon in source language, we can now continue with the translation task.

7.2.1 Bilingual Dictionary

Bilingual dictionaries can provide accurate translations but only for a few words because these dictionaries usually list word stem which yields a low-coverage problem. On the positive side such dictionaries often employ part-of-speech tags, so we might obtain different translations of the same word depending on its part-of-speech. To be able to use this information, however, we also need part-of-speech tagger in target language.

We use a bilingual English-Turkish dictionary[5] which contains 127157 entries. All entries in the dictionary have part-of-speech tags but since we do not have access to a part-of-speech

[5] http://www.fen.bilkent.edu.tr/~aykutlu/sozluk.txt

tagger in Turkish we discard this information when we translate the lexicon. We preserve, however, all the translation candidates for a word since one word in English can match with multiple number of Turkish words due to a semantic relationship (e.g. synonym) or simply bearing a secondary meaning of the same word.

Recall that we compiled SWN-weighted lexicon which consists of 38149 entry, from which we have found only 15611 entry has a translation listed in the dictionary which means more than half (59.1%) of the original lexicon could not be translated. Moreover the bilingual dictionary has no precedence among translation candidates, it is sorted alphabetically. Since we have a low hit rate and no information to obtain a dominant translation for a word, we decide to include all candidates for a word, therefore 15611 entry in the original lexicon yield 68028 translations (4.36 per entry). After compiling a raw translated lexicon, we first remove duplicate translations (42247 out of 68028) while calculating simple averages to assess a new polarity score to a word which has duplicate entries in the lexicon. Also, 53 words in translated lexicon have lost their subjectivity scores since the polarity scores of English words which translates to the same Turkish word cancel each other. Finally we end up with a sentiment lexiconwhich consists of 12231 positive and 13455 negative Turkish words. We refer to this lexicon as DICT in the remaining part.

7.2.2 Machine Translation

[25] uses Google Translate[6] for all machine translation tasks. In this experiment we use both Google Translate and Yandex Translator[7] to constrcut benchmark lexicons. Later when we build a lexicon with context-based translations on top of these benchmarks and we will only use Yandex Translator to translate the automatically created context due to the size limitation of translation engines.

After translating all the 38149 entries from SWN-weighted lexicon, we used a Turkish NLP library -Zemberek[8] to remove remaining English entries as the translation engines are not able to translate them all. Google Translate was able to translate 68% of the lexicon (25939 entry), whereas Yandex Translator could only translate 53.4% of the lexicon (20375 entry). These statistics also depend on the accuracy of Zemberek parser when it distinguishes a Turkish entry from a non-Turkish one, but our manual inspection shows that it is quite successful in this task. We use the same method (simple averages) to remove duplicate and neutralized entries as we did in Section 7.2.1. After all the processing, we compile a sentiment lexicon which consists of 8414 (6964) positive and 9132 (7458) negative unique entries by using Google (Yandex) Translator. Finally for the benchmark machine translated Turkish lexicon (MT-combined) we combine two lexicons (MT-Google) and (MT-Yandex), and after removing duplicate entries MT-combined lexicon ends up with 11118 positive and 12046 negative unique entries. We refer to this benchmark lexicon as MT in short for the remaining part. Figure 7.3 summarizes results provided in this section.

Context Based Approach

In [25], Meng et al. proposes a number of methods to prompt machine translation engines to return alternative translations for the same English word: Collocation, Coordinated phrase,

[6] http://translate.google.com/

[7] https://translate.yandex.com/

[8] http://zemberek.googlecode.com/svn/trunk/

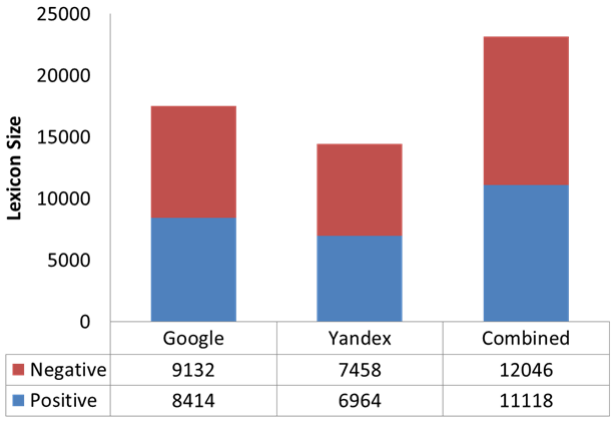


Figure 7.3: Number of sentiment words in MT lexicons

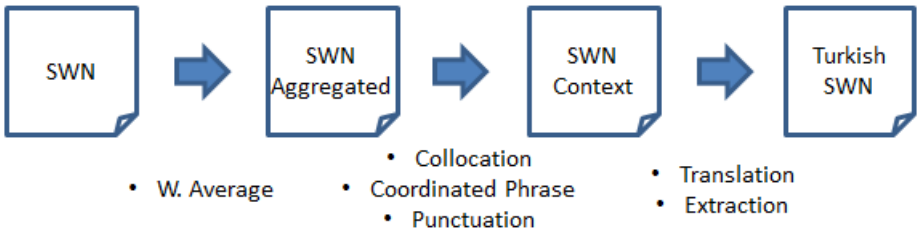


Figure 7.4: Context based translation of SentiWordNet

and Punctuation. We also use these methods with minor modifications on some time. We illustrate the work flow of creating Turkish SentiWordNet using context-based approach in Figure 7.4

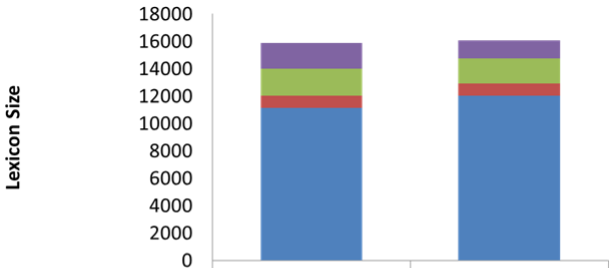
Collocation

This method suggests that providing a collocation along with the word in interest helps the translation engine to pick out more accurate translations. Following [25] we also use Microsoft N-gram Services [9], an online N-gram corpus built from Web documents. We choose the "bing-body/apr10/2" language model for this experiment which means that it is a bi-gram model compiled from the body of Web documents in April 2010.

Given each word w_1 in SWN-weighted lexicon, we use this model to generate up to the 20 most frequent bi-grams w_1w_2. After obtaining all these bigrams our SWN-weighted lexicon increases to 717651 entries and we use Yandex Translator to translate them to Turkish. After performing the translation we have to extract the word of our interest from its bi-gram form but it is not a straightforward task because of the differences in grammer of English and Turkish. We use a heuristic to perform this extraction to eliminate at least some of the known bigram translations using the translation of the collocation. This heuristic works as follows: besides with the whole bi-gram w_1w_2 we also fed the collocation w_2 to the translation engine and if the translation of the w_2 appears in the translation of w_1w_2 we remove the former from the latter to extract only the translation of w_1.

For example, we have the word "wooded" in the original SWN-weighted lexicon, and when

[9] http://web-ngram.research.microsoft.com/info/



	Positive	Negative
Bigram Web Search	1892	1293
Punctuation	1942	1859
Coordinated Phrases	916	863
MT	11118	12046

Figure 7.5: Number of sentiment words in CONTEXT lexicon

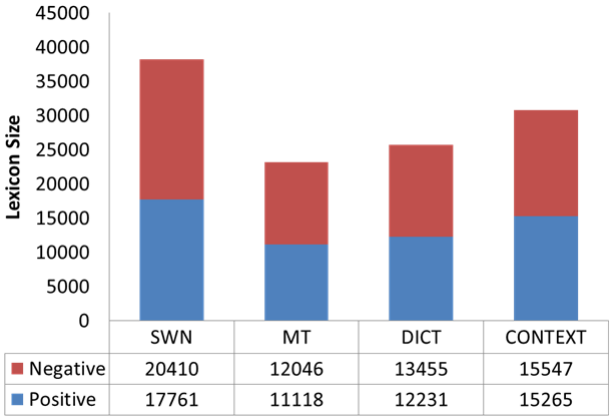
we generate its most frequent bi-grams one of them is "wooded areas". Yandex Translator translates it Turkish as "ormanlık alanları", while it translates the word "areas" as "alanları" so we have a common word "alanları" in both. After removing it from "ormanlık alanları" we obtain the Turkish word "ormanlık" as a translation of "wooded" and it is indeed an alternative translation for "wooded" because if we try to translate the word alone it returns the Turkish word "ağaçlık". In this example we gain one more extra word for our Turkish lexicon with proposed collocation method, however, we already have this alternative translation from the Google Translate. Thus part of the improvement from the benchmark MT lexicon reported in [25] would be lower when we combine two different machine translation engines.

After completing bi-gram extraction we remove non-Turkish, duplicate, neutralized entries and the phrases with more than two words (they are merely combinations of shorter phrases and often meaningless) in respective order. This returns 38210 positive and 37549 negative entries but when we manually inspect them we still spot a lot of noise. To keep only the most accurate translations we decide to look for search appearances of each new entry and preserve only top 10% as to be added to our sentiment lexicon. We use Microsoft Bing [10] for this task, and sort entries in decreasing order with respect to their search appearances. Then we manually remove some of the top hits because they are not Turkish i.e. the Turkish NLP library - Zemberek was not able to detect them as non-Turkish. Still there is an inherit ambiguity between some English and Turkish words. For example, the word "define" cannot be distinguished as English because it also used in Turkish with a different meaning. We leave to resolve these ambiguities for now because it requires us to go through all the candidates which conflict our automatic construction paradigm. At the end of all process we finally have 7359 candidates but only 1892 positive and 1293 negative entries are new i.e. does not appear in MT lexicon.

Coordinated phrase

We combine two English words that have the same Turkish translations to decrease the likelihood of obtaining same translations from the translation engine. We calculate that 4136 same Turkish translations obtained from multiple English words. Following example shows a number of English words with their sentiment orientations, and the translation engine

[10] http://www.bing.com/



	SWN	MT	DICT	CONTEXT
Negative	20410	12046	13455	15547
Positive	17761	11118	12231	15265

Figure 7.6: Comparison of lexicons in terms of size

translates them to the same Turkish word "zararlı"

zararlı : ('injurious', '-0.250'), ('deleterious', '-0.250'), ('prejudicial', '-0.375'), ('harmful', '-0.625'), ('pernicious', '-0.594'), ('baneful', '-0.594'), ('maleficent', '-0.750'), ('detrimental', '-0.750'), ('malefic', '0.125'), ('noxious', '-0.250')

Out of these 4136 Turkish translations we construct 19653 coordinated phrases. Below we put some example coordinated phrases constructed from the Turkish word "etkili":

effectual and influential = etkili ve nüfuzlu

effectual and effective = etkili ve etkili

effectual and take effect = güçlü ve etkili olmasi

As seen in the first example phrase we obtain the Turkish word "nüfuzlu" as an alternative translation for "etkili". More interestingly the first English word in the third phrase "effectual" translates to a Turkish word "güçlü" which is also different than its dominant translation. As a result, after extracting new translations from coordinated phrases and removing phrases with more than two words, we gained 916 positive and 863 negative new entries on top of the MT lexicon.

Punctuation

We place a punctuation mark (period) at the end of each English word as this method may effect the translation engines as it limits the possible parts-of-speech of the translations. Indeed we obtain 1942 positive and 1859 negative new entries on top of MT lexicon using this method. Figure 7.5 shows what each method adds up on top MT lexicon in terms of lexicon size.

7.3 Results & Discussion

In this section we assess the quality of translated lexicons by looking at two metrics: coverage of most frequent words, and accuracy of sentiment orientation.

7.3.1 Coverage of most frequent words

A list[11] of most frequently occurring words in Turkish was compiled in Princeton University. We use this list to assess how well our lexicons convey the most frequent words in modern Turkish. We do not expect a high coverage score since this list does not necessarily include only sentiment words, it has a lot noun with neutral meaning indeed. Hence the comparison in coverage might not give us a hint of their sentiment classification performance but on the other hand it gives us a good indication whether the methods that we use to build CONTEXT lexicon are effective in capturing some important frequent words that we might have missed otherwise. We list coverage results in Table 7.2. As seen in the table CONTEXT lexicon has indeed a better coverage but still less than DICT. Part of the reason that DICT lexicon has relatively high coverage is both bilingual dictionary and the list of frequent words have only word stem which increases the likelihood of possible match. Here we could use Zemberek, an NLP library for Turkish to obtain word stem of the entries from machine translation but initial tests with its lemmatizer does not return reliable results so we discard this option.

Table 7.2: Frequent words coverage

	DICT	MT	CONTEXT
Coverage	76.5	62.7	73.4

7.3.2 Accuracy of sentiment orientation

We manually annotate a small sized random samples from each Turkish lexicon to assess how accurate they are in terms of sentiment orientation of their entries. Two native Turkish speaker performed the annotation tasks, and disagreements resolved through discussion. Without considering non-polar entries all three lexicons have high accuracy, however they all suffer huge number of non-polar entries and overall results turns out to be poor. We list the summary of results in Table 7.3

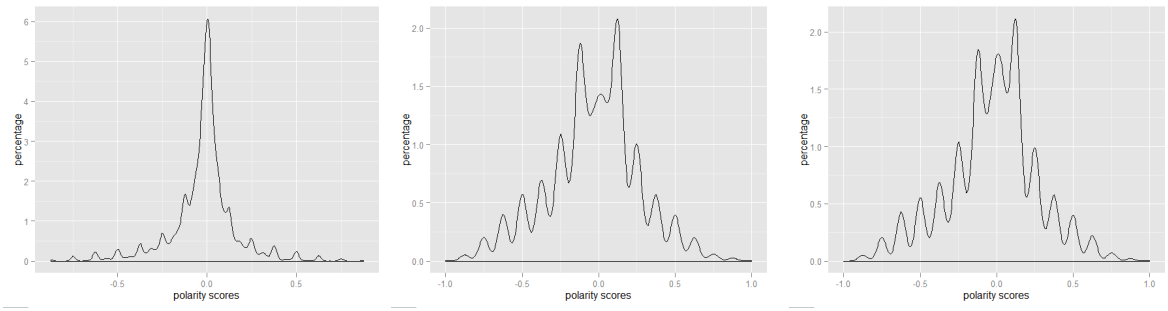
Table 7.3: Accuracy in sentiment orientation

	DICT	MT	CONTEXT
w.o non-polar	95	95	96
overall	30.83	25	31.37

7.3.3 Effectiveness in Polarity Detection

Polarity detection is one the of main application areas for sentiment lexicons. Not only they can form a basis for rule-based classifiers, but they are also useful as a seed to build more complex features for corpus-based classifiers. Thus we also want to evaluate the lexicons on detecting polarity of a document. For this task we use a multi-domain product reviews dataset in Turkish which was first introduced in [9]. This dataset consists user reviews from books, dvd, electronics, and kitchen appliances departments (See Chapter 4 for more details about the dataset). Similar to the rule-based classifier we used to classify English reviews in

[11] http://www.turkishlanguage.co.uk/freqvocab.htm



(a) DICT (b) MT (c) CONTEXT

Figure 7.7: Desnsity graphs for translated lexicons from SWN-weighted

Section 7.2.2, we followed the same rule to construct a rule-based classifier in Turkish: it gives a polarity score to each review by adding up polarity scores of the words appear in it.

Table 7.4, 7.5, 7.6, 7.7 show the results of the rule-based classifiers which use unigram feature extraction and MT, DICT, and CONTEXT lexicons respectively. We also experimented with bigram features but since the results were similar to those from unigram, we only present the results of unigram features here. We show the independently measured precision and recall for positive and negative reviews, and the total accuracy. As seen from the tables, among the three lexicons DICT outperforms the other two in every domain both in terms of total accuracy. This result contradicts the result presented in [25] for Chinese as they report that the Chinese lexicon created using a bilingual dictionary underperforms in sentiment classification comparing to MT, and CONTEXT lexicons. Furthermore our result does not support their claim that bilingual dictionaries are not effective for adapting resources cross-lingually. Hovewer we have to note that even though DICT performs the best in our experiment setting, still its accuracy score is on the low side when we compare these results with corpus-based classification results presented in [9] which reports around 80% accuracy for all domains. This shows that comparing to translated lexicons corpus-based methods are better suited in classifiacation of reviews by considerable extent. Similarly in [26], the two rule-based subjectivity classifiers (F-measure 43.66% and 47.93%) implemented using a lexicon translated to Romanian are also compared with a corpus-based classifier (F-measure 67.85%) and their results are compatible to ones we obtained in this experiment.

Since our rule-based classifiers perform poorly we also try to combine them with baseline classifiers to see if they have any potential to improve classification performance. We obtain probability estimates from baseline classifiers for each test instance and when they read low, we delegate the task to our rule based classifier and use its results. Each machine learning algorithm produces different probability estimate distributions due to their implementation details, and even for our simple heuristic to combine two classifiers we need to know which probability levels we should regard as low. Naive Bayes classifier is tend to produce probabilities which converge to each extreme [-1,+1] whereas logistic regression classifiers (Maximum Entropy) produce more even distribution. Thus we only use the results from NB classifier if it's probability estimate reads above 0.85 or below -0.85, and for ME classifier we set these numbers as 0.5 and -0.5. We report no significant improvement in classification results mostly since our rule based classifier also work with same feature set (unigrams) but with much lower accuracy which makes the combination task obsolete.

Table 7.4: Sentiment classification comparisons (Books)

Methods	Positive		Negative		Total
	Precision	Recall	Precision	Recall	Accuracy
DICT	57.4	90.14	76.7	32.85	61.5
MT	54.49	80.85	62.49	32.28	56.57
CONTEXT	54.49	84.28	64.84	29.43	56.85
NB	81.56	82.14	83.49	79.14	80.64
ME	80.38	78.85	80.4	78.28	78.57
NB+DICT	80.61	84.14	84.74	77.57	80.86
ME+DICT	78.86	80.14	81.45	76.71	78.42

Table 7.5: Sentiment classification comparisons (DVD)

Methods	Positive		Negative		Total
	Precision	Recall	Precision	Recall	Accuracy
DICT	57.55	84.71	71.27	37.42	61.07
MT	54.71	80.14	63.02	33.57	56.85
CONTEXT	54.35	83.14	64.02	30.0	56.57
NB	77.19	71.28	73.52	78.0	74.64
ME	75.49	72.71	74.13	75.57	74.14
NB+DICT	75.85	72.14	73.49	75.57	73.85
ME+DICT	73.65	71.85	72.65	73.14	72.5

The results also show that our rule based classifier is poor in detecting negative reviews correctly in case of all three lexicons. We tried to experiment with different thresholds to determine if a review as positive or negative, however, the change in overall accuracy of the classifiers remains low. Thus we leave the threshold at the zero. To check if the lexicons have any bias towards to positive side, we also construct density graphs for DICT, MT, and CONTEXT lexicons in Figure 7.7. However the density graphs do not spot any abnormal distribution or bias towards one side. While density distribution of MT and CONTEXT lexicons are similar, we see that the DICT lexicon has vast majority of sentiment phrases whose polarity scores close to zero. Considering its relative success over other two lexicons in polarity detection, it shows a reverse correlation between the percentage of phrases which are not bearing any sentiment but misclassified after the translation and the accuracy of classifiers which use these phrases.

Table 7.6: Sentiment classification comparisons (Electronics)

Methods	Positive		Negative		Total
	Precision	Recall	Precision	Recall	Accuracy
DICT	56.76	84.71	69.14	35.14	59.92
MT	56.92	65.71	59.92	49.85	57.78
CONTEXT	56.68	72.57	60.86	43.85	58.21
NB	77.96	88.71	85.44	66.42	77.57
ME	78.23	78.28	77.0	71.57	74.92
NB+DICT	77.28	87.57	82.67	66.42	77.0
ME+DICT	77.77	79.14	76.96	71.42	75.28

Table 7.7: Sentiment classification comparisons (Kitchen)

Methods	Positive		Negative		Total
	Precision	Recall	Precision	Recall	Accuracy
DICT	54.07	86.42	66.29	26.57	56.5
MT	55.21	66.0	57.71	46.42	56.21
CONTEXT	54.92	74.14	60.21	39.14	56.64
NB	71.26	81.57	77.92	65.28	73.42
ME	72.27	70.28	70.91	72.14	71.21
NB+DICT	69.45	81.85	77.87	63.0	72.43
ME+DICT	71.1	72.14	71.49	69.86	71.0

Chapter 8

Conclusions

8.1 Summary of results

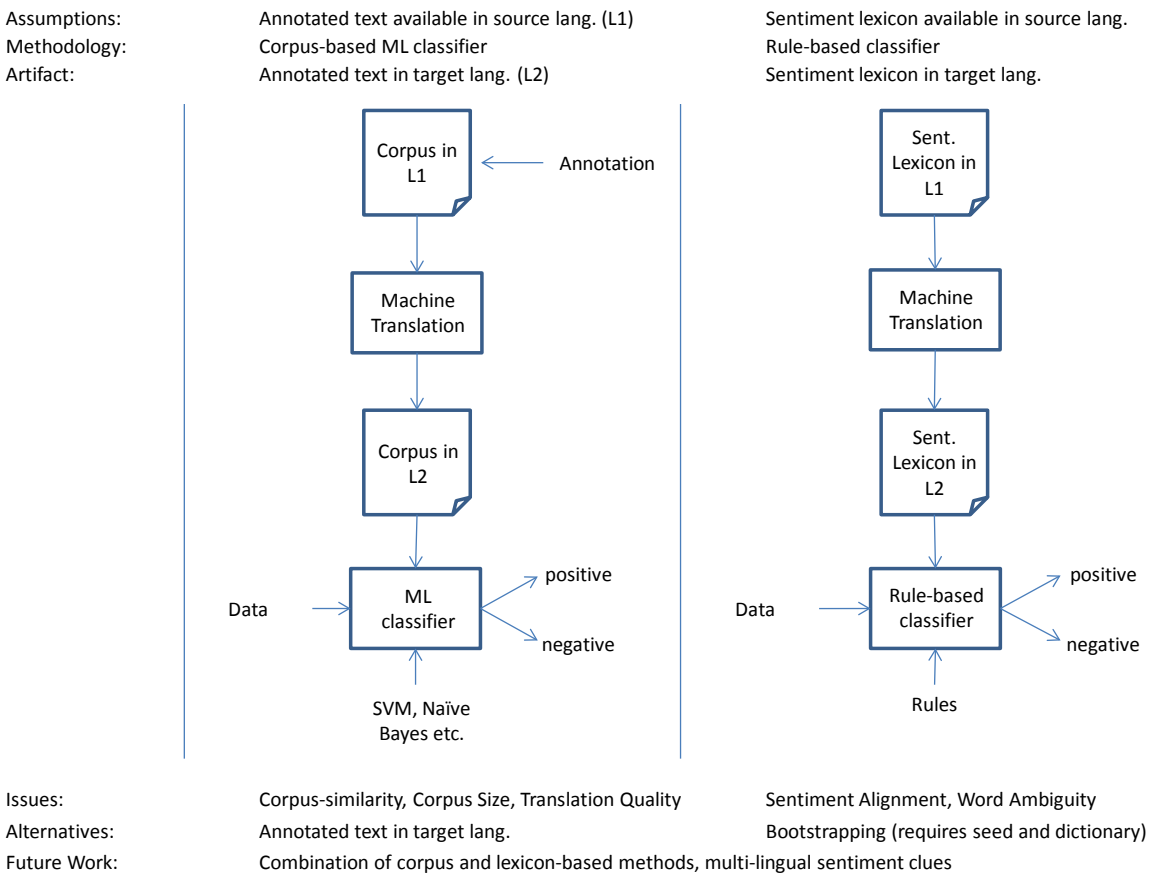


Figure 8.1: Summary of experiment setups

This thesis is about cross-lingual projections to leverage existing resources from English using machine translation. Besides a purely scientific interest, our interest in learning cross-lingual sentiment classifier is a pragmatic one, motivated by the inherent trade-off between prediction performance and development cost. Fully labeled data is typically costly and time-

consuming to produce and requires specialist expertise, but when available typically allows more accurate prediction. Unlabeled data, on the other hand, is often available at practically zero marginal cost. In Chapter 5 we provide empirical evidences that shows automatically created labeled data allow us to strike a balance between annotation cost and prediction accuracy.

In Chapter 6, we describe that amount of artificial noise added by machine translation services does not hinder classifiers much in polarity detection task. However, it is important to distinguish the effect of machine translation from the effect of merging different cross-lingual data sources and that like in case of transfer learning we may need to search for ways to account for cross-lingual data distribution differences.

In Chapter 7, we present three methods to translate a sentiment lexicon to another language. The first method is to use a bilingual dictionary, and the other two rely on machine translation capabilities where we also try to increase the number of candidates for a translation by generating a context for the translation engine. We conduct an experiment which we obtain comparable results with previously published studies in [26, 25]. We pick SentiWordNet as a source lexicon and translate it to Turkish, a language with limited resources and tools for sentiment analysis. We follow a number of approaches presented in [25] to generate the context for the translation engine such as looking for frequent collocations of a word, combining words that have the same translation, and make use of punctuation to limit the possible part-of-speech of a translation.

We evaluate the lexicons we built as part of the experiment in terms of accuracy in expressing the correct sentiment in phrase level, coverage of most frequently used words, and their ability to detect polarity of documents when they are used as base dictionaries for rule-based classifiers. We highlight through our experiment results that assessing the correct sentiment of a translated word has inherently difficult because of cultural and linguistic differences even though the translation itself is correct. Moreover we still could not translate a significant portion of the SentiWordNet (32% for Google, 46.6% for Yandex) to Turkish using machine translation engines. We also lost any clue to resolve word ambiguity in the translation which makes it even harder to detect any sarcasm or figurative meanings in documents. Automated methods are tend to produce a noticeable amount of noise that we spot by manual inspection. Most of this noise comes from a translation that has no use in target language i.e. meaningless phrases which are just mere combination of shorter ones. On another side mark, all three lexicons are poor in detecting sentiment of product reviews while the lexicon translated by a bilingual dictionary outperforms the others.

As a consequence, although the methods presented in [25] is also successful to expand lexicon size in Turkish, we cannot raise a claim to support some of their results as for CONTEXT lexicon they show a significant improvement from DICT and MT lexicons with respect to the accuracy of classifying sentiment in Chinese.

8.2 Contributions

We can list our contributions by giving a brief answer to four research questions formulated in Chapter 1. Answer to the first question highlights overall view whereas the remaining three related to several scenarios that we can encounter dealing with a new language.

- **How useful the resources created by machine translation in sentiment analysis? Is it a viable alternative to time-consuming manual annotation task?**

In Chapters 5,6 and 7, we explored several approaches to employ machine translation in cross-lingual sentiment analysis. Following the previous research in the literature our approaches can be divided into two parts: corpus-based, and lexicon-based. In Chapter 5, we show that translating an annotated corpus to another language by preserving the actual annotations results at most 3% drop in major accuracy of general-purpose (Naive Bayes, SVM) classifiers.

In Section 6.2, we study the co-training to learn polar text in target language as it was proposed in [40]. Co-training method stands out a viable alternative to in-language classifiers as the results show that it can improve the polarity classification of text in target language if a raw data is already available in target language as well.

However, in Chapter 7, when we build target language classifier by translating an existing sentiment lexicon from English, we realize that translation methods fail to capture inherit divergence between languages with respect to expression of sentiments. Results also show that such classifiers poorly perform in polarity detection comparing to corpus-based counter sides.

- **Assuming we have access to additional labeled data from another language, how can we make use of it to improve the classification performance?**

In section 6.1, we observe that in the presence of in-language training data, adding additional data from other languages results in no improvement to the classifier, and indeed it may harm the performance. Introducing data from another language resembles to the problem of using data from another domain as they both need a method to resolve dissimilarity between different sources.

- **Assuming we have access to additional raw data from another language, how can we make use of it to improve the classification performance?**

This question is the main motivation of co-training experiments. Experiment results reveal that the claimed benefits of co-training approach in [40] is likely the effect of semi-supervised learning where a classifier can learn more from raw data in each iteration. Although the combined classifier outperforms the language-specific classifiers, difference among them is less than 1% base-points in accuracy. Thus we state that raw data from another language can help in classification of text in the language of interest if the additional data is used to learn new features as in the semi-supervised learning.

Recall that in answering previous question we state additional labeled data provides no help if they are directly merged to existing training data in another language. Following the experiment with raw data, it confirms that the effort to annotate text in a language is only feasible if one plans to build in-language sentiment classifier as we cannot directly benefit from those annotations in cross-lingual sentiment analysis, and obtaining raw data is far more easier than annotated resources.

- **Assuming we have no access to any training data in our language of interest, then what are the ways to translate a sentiment lexicon to perform sentiment classification in that language?**

In Chapter 7, we use these two approaches to translate a sentiment lexicon to Turkish and determine the translation quality with respect to coverage of most frequently used words in Turkish, accuracy of sentiment orientation, and effectiveness in polarity detection. Results show that bilingual dictionary outperforms machine translation benchmark in all three metrics. Even when we improve machine translation accuracy by generating context for translation, we obtain an experimental evidence of bilingual dictionary method to be more accurate.

8.3 Future Work

Our experiments in Chapter 6 show that naive ways of introducing new sources from other languages causes cross-domain dissimilarity issues. This indicates that existing approaches applicable to cross-domain sentiment classification, e.g. [13] and further advancement in this direction might be fruitful for cross-lingual sentiment analysis too. This is one of the directions of our future work.

We have studied how machine translation affects the performance of the general purpose classification techniques. In the future work we plan to consider also techniques specific to sentiment classification like e.g. a rule-based approach to polarity detection [38].

In future we plan to elaborate the experiments by using different source language lexicons to analyze how much the inconsistency in sentiment orientation of a word depends on the actual lexicon in source language but not on erroneous translation. We will also try to explore a more robust schema to assess a sentiment score to a translated word independent from its source language score, for this task part of the lexicon can be used to propagate domain-specific sentiment words if an annotated domain corpus is available.

Bibliography

[1] A. Balahur and A. Montoyo. Semantic approaches to fine and coarse-grained feature-based opinion mining. In *Proceedings of the 14th international conference on Applications of Natural Language to Information Systems*, NLDB'09, pages 142–153, Berlin, Heidelberg, 2010. Springer-Verlag. 7

[2] A. Balahur and M. Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 2013 (In press). 11, 15, 16

[3] C. Banea, R. Mihalcea, and J. Wiebe. Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 28–36, 2010. 19

[4] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, 2008. 3, 10

[5] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL'07, pages 187–205, 2007. 7, 13, 29

[6] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of RANLP'2009*, pages 50–54, 2009. 11

[7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, 2013. 19

[8] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 355–362, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. 6

[9] E. Demirtas and M. Pechenizkiy. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 9:1–9:8, New York, NY, USA, 2013. ACM. 34, 35

[10] K. Denecke. Are sentiwordnet scores suited for multi-domain sentiment classification? In *ICDIM*, pages 33–38, 2009. 27, 29

[11] K. Duh, A. Fujino, and M. Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 429–433, 2011. 11

[12] A. Esuli, S. Baccianella, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10*, 2010. 26

[13] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 513–520, 2011. 41

[14] A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 45–52, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 6

[15] A. Hassan, A. Abu-Jbara, R. Jha, and D. Radev. Identifying the semantic orientation of foreign words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 592–597, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 27

[16] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the ACL*, pages 174–181, 1997. 7, 26

[17] R. Johansson and A. Moschitti. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509, 2013. 6

[18] S.-M. Kim and E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 6

[19] S.-M. Kim and E. Hovy. Identifying and analyzing judgment opinions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 200–207, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 6, 27

[20] L. Lee. A matter of opinion: Sentiment analysis and business intelligence (position paper). *CoRR*, abs/cs/0504022, 2005. 5

[21] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012. 5

[22] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 415–463. Springer US, 2012. 26

[23] A. maria Popescu and O. Etzioni. Extracting product features and opinions from reviews. pages 339–346, 2005. 7

[24] R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007. 7

[25] X. Meng, F. Wei, G. Xu, L. Zhang, X. Liu, M. Zhou, and H. Wang. Lost in translations? building sentiment lexicons using context based machine translation. In *COLING (Posters)*, pages 829–838, 2012. 3, 27, 30, 31, 32, 35, 39

[26] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, 2007. 10, 27, 35, 39

[27] S. min Kim. Determining the sentiment of opinions. In *In Proceedings of COLING*, pages 1367–1373, 2004. 7

[28] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *In Proceedings of the ACL*, pages 271–278, 2004. 6, 7

[29] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, pages 115–124, 2005. 6, 13

[30] B. Pang and L. Lee. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval 2(1-2)*, 2008. 5

[31] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP02*, 2002. 6, 7

[32] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 25–32, 2003. 7, 26

[33] M. Sahlgren, J. Karlgren, and G. Eriksson. Sics: Valence annotation based on seeds in word space. In *In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007*, pages 296–299, 2007. 6

[34] J. Savoy and L. Dolamic. How effective is google's translation service in search? *Commun. ACM*, 52(10):139–143, Oct. 2009. 11

[35] C. Scheible. Sentiment translation through lexicon induction. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 25–30, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 27

[36] J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vázquez, and V. Zavarella. Creating sentiment dictionaries via triangulation. *Decis. Support Syst.*, 53(4):689–694, 2012. 11

[37] P. J. Stone and E. B. Hunt. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, AFIPS '63 (Spring), pages 241–256, 1963. 5

[38] E. Tromp and M. Pechenizkiy. Rbem: A rule based approach to polarity detection. In *Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM@KDD'13)*. ACM, 2013. 41

[39] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *CoRR*, cs.LG/0212032, 2002. 6, 7

[40] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL'09, pages 235–243, 2009. 3, 19, 23, 25, 40

[41] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pages 486–497, Berlin, Heidelberg, 2005. Springer-Verlag. 7, 8

[42] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, Sept. 2004. 5, 7

[43] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, 2005. 7, 10

[44] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 761–767. AAAI Press, 2004. 7

[45] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 6