EXTRAPOLATING SUBJECTIVITY RESEARCH TO OTHER LANGUAGES

Carmen Banea

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2013

APPROVED:

Rada Mihalcea, Major Professor
Janyce Wiebe, Committee Member
Paul Tarau, Committee Member
Jiangping Chen, Committee Member
Barrett Bryant, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
Mark Wardell, Dean of the Toulouse Graduate School

Banea, Carmen. Extrapolating Subjectivity Research to Other Languages. Doctor of Philosophy (Computer Science and Engineering), May 2013, 125 pp., 23 tables, 26 illustrations, 88 numbered references.

Socrates articulated it best, "Speak, so I may see you." Indeed, language represents an invisible probe into the mind. It is the medium through which we express our deepest thoughts, our aspirations, our views, our feelings, our inner reality. From the beginning of artificial intelligence, researchers have sought to impart human like understanding to machines. As much of our language represents a form of self expression, capturing thoughts, beliefs, evaluations, opinions, and emotions which are not available for scrutiny by an outside observer, in the field of natural language, research involving these aspects has crystallized under the name of subjectivity and sentiment analysis. While subjectivity classification labels text as either subjective or objective, sentiment classification further divides subjective text into either positive, negative or neutral.

In this thesis, I investigate techniques of generating tools and resources for subjectivity analysis that do not rely on an existing natural language processing infrastructure in a given language. This constraint is motivated by the fact that the vast majority of human languages are scarce from an electronic point of view: they lack basic tools such as part-of-speech taggers, parsers, or basic resources such as electronic text, annotated corpora or lexica. This severely limits the implementation of techniques on par with those developed for English, and by applying methods that are lighter in the usage of text processing infrastructure, we are able to conduct multilingual subjectivity research in these languages as well. Since my aim is also to minimize the amount of manual work required to develop lexica or corpora in these languages, the techniques proposed employ a lever approach, where English often acts as the donor

language (the fulcrum in a lever) and allows through a relatively minimal amount of effort to establish preliminary subjectivity research in a target language.

ACKNOWLEDGMENTS

Despite studying subjectivity in natural language, I have always felt at a loss of words when expressing gratitude toward the wonderful people that have made a difference in my life. I hope that they can read between the lines and know that I could not be who I am today without their shoulder to lean on during harder times, and their unwavering love and support despite my failures; and for this, I am thanking you from the bottom of my heart.

I would love to thank Rada, the best advisor that a doctoral student could ever dream of. She is one of the warmest, kindest hearted people I know, who taught me to take everything with a good dose of good will and humor. Her mentoring goes beyond just research. She supported me and stood by me in the most difficult times in my life. She is that fabled advisor that one aspires to be, and I consider myself so fortunate to have been her student.

My husband, Samer, has brightened my days ever since we met, and he has been an incredible motivational force throughout my life. He has always mastered the skill of making me laugh even when I felt that all the ships were sinking, of master minding new foods when I could not eat, of holding my hand when my feet were slipping. Thank you!

This dissertation is dedicated to my parents, who have sacrificed everything in their life for my sister and I, and who have not taken a single decision without bearing us in mind. I will never be able to comprehend the amount of love, care and selflessness they have shown me, and I find nature so merciless to never allow children to give back even a fraction of what we have received. Many thanks to my sister, Raluca, who has always believed in me and to my grandmother.

To my numerous amazing friends, spread all over the world, who have enriched my life, thank you!

Last but not least, I would like to thank my PhD committee for spending their valuable time and effort to guide me through my research and dissertation.

Thanks Blondie (my cat) for looking over my dissertation and Simba (my dog) for being a good and patient listener! You are the sweetest souls that keep me company day and night!

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels text as either subjective or objective, sentiment classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral.

To date, a large number of text processing applications have already used techniques for automatic sentiment and subjectivity analysis, including automatic expressive text-to-speech synthesis [2], tracking sentiment timelines in on-line forums and news [41, 4], analysis of political debates [67, 14], and mining opinions from product reviews [25]. In many natural language processing tasks, subjectivity and sentiment classification have been used as a first phase filtering to generate more viable data. Research that benefitted from this additional layering ranges from question answering [85], to conversation summarization [13], to text semantic analysis [76, 19], and word sense disambiguation [1].

1.1. Problem Definition

Much of the research work to date on sentiment and subjectivity analysis has been applied to English, but work on other languages is growing, including Japanese [35, 64, 66, 30], Chinese [26, 68, 87], German [34], and Romanian [43, 10]. In addition, several participants in the Chinese and Japanese Opinion Extraction tasks of NTCIR-6 [32] performed subjectivity and sentiment analysis in languages other than English.

As only 27% of Internet users speak English,[1] the construction of resources and tools for subjectivity and sentiment analysis in languages other than English is a growing need. Figure 1.1 shows the growth in electronic text available on the Internet over the last decade. While English has experienced a moderate growth of 301.40%, languages such as Arabic, Russian and Chinese

[1] www.internetworldstats.com/stats.htm, Oct 11, 2011

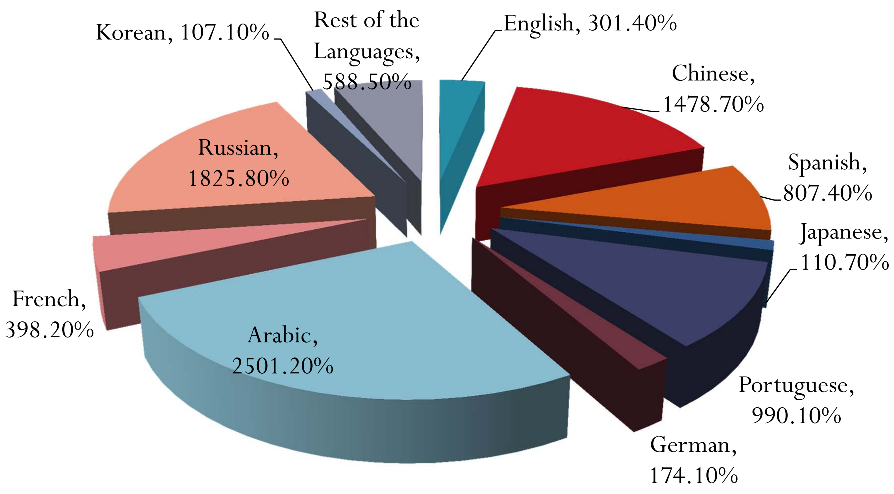


FIGURE 1.1. Growth in Internet languages over the last decade[2].

exhibited a four digit percentage growth, with Portuguese and Spanish lagging closely behind, at 990% and 807%, respectively.

1.2. Proposed Solution

In this work, I investigate techniques of generating tools and resources for subjectivity analysis that do not rely on an existing natural language processing infrastructure in a target language. This constraint is motivated by the fact that the vast majority of human languages are scarce from an electronic point of view: they lack basic tools such as part-of-speech taggers, parsers, or basic resources such as electronic text, annotated corpora or lexica. This severely limits the implementation of techniques on par with those developed for English, and by applying methods that are lighter in the usage of text processing infrastructure, we are able to conduct multilingual subjectivity research in these languages as well. Since my aim is also to minimize the amount of manual work required to develop lexica or corpora in these languages, the techniques presented here employ a lever approach, where English often acts as the donor language (the fulcrum in a lever) and allows through a relatively minimal amount of effort to establish preliminary subjectivity research in a target language.

While much recent work in subjectivity analysis focuses on *sentiment*, I opt to focus on recognizing subjectivity in general, for two reasons.

First, even when sentiment is the desired focus, researchers in sentiment analysis have shown that a two-stage approach is often beneficial, in which subjective instances are distinguished from objective ones, and then the subjective instances are further classified according to polarity [85, 48, 83, 34]. In fact, the problem of distinguishing subjective versus objective instances has often proved to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification. This is reported in studies of manual annotation of phrases [66], recognizing contextual polarity of expressions [83], and sentiment tagging of words and word senses [3, 19].

Second, a natural language processing (NLP) application may seek a wide range of types of subjectivity attributed to a person, such as their motivations, thoughts, and speculations, in addition to their positive and negative sentiments. For instance, the opinion tracking system Lydia [41] gives separate ratings for subjectivity and sentiment. These can be detected with subjectivity analysis but not by a method focused only on sentiment.

1.3. Contributions

My contributions encompass multiple aspects, as I take into consideration the various granularities at which subjectivity can be expressed in text, as well as various methods that allow for subjectivity research to be extended to numerous languages, most importantly those with scarce electronic resources. Furthermore, some of the methods allow for subjectivity detection to be improved in English as well. Below is an enumeration of several questions that this work seeks to answer:

i. Can subjectivity research be carried out in languages other than English without requiring in language resources?

   In Chapters 4, 5 and 6 I explore several approaches to conducting subjectivity research in languages other than English that require no subjectivity resources to be available in the target language. These approaches leverage the rich NLP infrastructure available for English and seek to port the necessary information to the target language. The recipient language thus acquires subjectivity lexica and corpora, as well as subjectivity analysis tools, which set the basis for conducting sentiment and subjectivity research in the given language.

Subjectivity annotated lexicons. I focus on two potential paths to leverage on subjectivity annotated lexicons. One is based on attempting to automatically translate a source language lexicon into the target language by selecting the first sense from a bridging multilingual dictionary. The resources required by this experiment are an annotated subjectivity lexicon and a multilingual dictionary. The second one is based on selecting a small set of subjective seeds from the source language lexicon and manually translating them into the target language. This lexicon can be expanded by solely using material in the target language through a bootstrapping mechanism. This scenario requires the availability of a small set of subjective seeds, an electronic dictionary, and a raw corpus.

Manually annotated corpora for subjectivity. I propose leveraging on a corpus manually annotated for subjectivity at the sentence level. In order to create a version of the corpus in the target language, a statistical machine translation engine is employed that allows transfer between the source and target language. Since the resulting corpus in the target language is not annotated, the human annotations from the source corpus are projected onto the machine translated corpus in the target language. The resulting annotated corpus can then be used to train a machine learning algorithm and therefore create an automatic subjectivity analysis tool in the target language. This experimental setup requires the availability of a manually annotated corpus for subjectivity in the target language, as well as a machine translation engine between the source and the target language.

Subjectivity annotation tools. This scenario explores the potential of using parallel text paired with an automatic tool for subjectivity analysis to generate subjectivity annotated text in the target language. We look at the differences between employing manual parallel text as well as parallel text generated automatically through machine translation. The annotated corpus developed in the target language will be able to act as a training data set for a machine learning algorithm. This setup requires either a manually translated parallel text, or data either in the source or target language paired with a machine translation engine.

ii. Is there a benefit for subjectivity research in considering several languages at the same time? Would a multilingual model of subjectivity be more robust when compared to a traditional monolingual approach?

In Chapters 4 and 6, I focus on setups that involve multilingual vectorial spaces and contrast their subjectivity modeling ability against more traditional monolingual spaces.

iii. Is subjectivity a language independent phenomenon, that is able to permeate language boundaries?

This research presents several manual annotation studies conducted at the sense and sentence level that are aimed at providing an answer to this question.

iv. Is a multilingual (or monolingual) dictionary sufficient in porting subjectivity lexica to a target language?

This facet is examined in Chapter 5, where I address the challenges encountered when using a multilingual dictionary. I also propose a monolingual word expansion method that is able to acquire a subjectivity lexicon with in-language resources.

v. Are machine translation systems able to transfer the subjective content from one language to another?

I explore this question by comparing the performance of subjectivity machine learning systems trained on human translated parallel text versus text resulted from a machine translation engine.

vi. Can we conduct multilingual subjectivity research at different granularities, be it the sense level, word level, or sentence level?

Chapters 4, 5 and 6 7 present methods for performing multilingual subjectivity research at different granularities.

vii. Are there additional markers of subjectivity that appear in some languages, but not in others? What would be some unique markers of subjectivity in Romanian?

Chapter 7 takes a quantitative and qualitative approach to multilingual subjectivity analysis first by analyzing the top features obtained as a result of feature selection and their correlation with human subjectivity judgments. It also identifies unique markers of subjectivity that are present in Romanian, such as verb conjugation, polite, formal and informal register, etc.

1.4. Thesis Outline

The thesis is organized as follows. Chapter 2 introduces the reader to a background on subjectivity analysis and its challenges, as well as some of its most notable applications. Chapter 3 covers the related work in this domain, starting with an overview of English related research, yet mostly focusing on multilingual approaches to subjectivity analysis. Chapters 4, 5, and 6 detail my proposed models in conducting multilingual subjectivity at the sense-level, word level, and sentence level, respectively. Each of these chapters include the proposed method, the experiments, the evaluations and the pertaining discussions. Prompted by the conclusions ensuing from these experiments, in Chapter 7, I conduct a qualitative analysis of language dependent factors that convey subjectivity by considering a case in point discussion of Romanian, which is explored by drawing a parallel with the way subjectivity is expressed in English. Finally, I conclude with Chapter 8, where I revisit the research questions and summarize how this thesis addressed them, as well as provide potential future work directions.

CHAPTER 2

BACKGROUND

2.1. Subjectivity

An important kind of information that is conveyed in many types of written or spoken discourse is the mental or emotional state of the writer, speaker, or some other entity referenced in the discourse. From early childhood we are exposed to stories, where the characters laugh, cry, inspire, amuse, strike fear, etc. These stories are aimed at teaching not only philosophical concepts of right and wrong, but they also aid us in inferring from discourse sentiments, evaluations, speculations, etc. Later on, during school years, as students become exposed to prose, poetry, theater, their ability to discern expressions of private text in language is further tuned. This prepares us for the adult life, where we are constantly exposed to subjective content, in the form of news articles, editorials, blogs, reviews, speeches, and television shows, and where our ability to discern factual information from the opinions overlaid by various sources becomes critical to understand events and interact with others. News articles, for example, often report the emotional response to a story along side the facts, and many times imprint a reader with a particular point of view. Editorials, reviews, weblogs, and political speeches also convey the opinions, beliefs, or intentions of the writer or speaker.

Wiebe et al. give us the general term *private state* to refer to this type of content [80], which groups together expressions of opinion, emotion, sentiment, evaluations, beliefs and speculations in natural language. Numerous researchers have taken upon the task of automatically identifying private states in discourse, which in the field of natural language processing is termed as subjectivity and sentiment analysis. While subjectivity analysis seeks to identify the private state being enunciated (subjective text), as well as its source (the entity holding the private state) and its target (the entity the private state concerns), sentiment analysis focuses on the polarity of the private state, whether it is positive, negative or neutral.

Let us consider the following excerpt from an article regarding the 2012 presidential campaign:[1]

> "Governor Romney apparently fears that the more he offers, the more our campaign will demand that he provide," Mr Messina wrote. (1)

This quote contains several private states that have different sources and targets. First of all the insertion of the adverb "apparently" by Mr. Messina, Barack Obama's presidential campaign manager in the 2012 election, indicates an assumption that he is making on behalf of the republican presidential candidate, Mitt Romney; its standalone polarity is neutral. The word plays a double role: first it maintains political correctness, as Messina can avoid committing himself to the ensuing statement, and second, it allows for a certain level of irony to be injected in the discourse, aiming to show that Romney has irrational fears. The next word "fears" indicates a negative emotion held by Romney regarding requests for additional income tax returns to be publicly released. Another marker of subjectivity is the verb "demand," as it signals both urgency and entitlement, and it has a negative connotation. Overall, the entire statement is subjective.

From Example 1, we notice that we may judge the subjectivity and polarity of words in and out of context. "Apparently" marks a speculation, with a neutral polarity, while "fear," directly signals a negative private state regardless of context; both these words are triggers of subjectivity. On the other hand, the subjectivity and polarity of "demand" is highly dependent on its surrounding context (see Table 1 for the various senses of "demand" as listed in WordNet 3.1 [45]), as some of its senses do not encode for a private state, and by extension for a polarity. The statement "the demand for methane has increased with the higher price of gas" is a purely objective statement, as demand is used as a noun, in its sense of economic supply and demand. In the case of "the judge demanded that he testifies in the trial," despite being used in the same verbal form as in the article excerpt, the sense is that of summoning to court, which is an objective usage.

Chen [15] mentions that several different dimensions may influence subjectivity in text. Among them, she identifies non-objectivity, uncertainty, vagueness, and ambiguity:

[1] `http://www.bbc.co.uk/news/world-us-canada-19299676`

Annotation	Definition & sample usage
Subj.	(n) demand (an urgent or peremptory request) — *his demands for attention were unceasing*
Obj.	(n) demand (the ability and desire to purchase goods and services) — *the automobile reduced the demand for buggywhips*; *the demand exceeded the supply*
Obj.	(n) requirement, demand (required activity) — *the requirements of his work affected his health*; *there were many demands on his time*
Subj.	(n) demand (the act of demanding) — *the kidnapper's exorbitant demands for money*
Obj.	(n) need, demand (a condition requiring relief) — *she satisfied his need for affection*; *God has no need of men to accomplish His work*; *there is a demand for jobs*
Subj.	(v) demand (request urgently and forcefully) — *The victim's family is demanding compensation*; *The boss demanded that he be fired immediately*; *She demanded to see the manager*
Obj.	(v) necessitate, ask, postulate, need, require, take, involve, call for, demand (require as useful, just, or proper) — *It takes nerve to do what she did*; *success usually requires hard work*; *This job asks a lot of patience and skill*; *This position demands a lot of personal sacrifice*; *This dinner calls for a spectacular dessert*; *This intervention does not postulate a patient's consent*
Obj.	(v) demand, exact (claim as due or just) — *The bank demanded payment of the loan*
Obj.	(v) demand (lay legal claim to)
Obj.	(v) demand (summon to court)
Subj.	(v) demand (ask to be informed of) — *I demand an explanation*

TABLE 2.1. A listing of the various senses of "demand" (as listed in WordNet 3.1) with their definition (or gloss) and usage samples (in italics) accompanied by their subjectivity annotation; the Subj. / Obj. labeling stands for subjective and objective, respectively.

- Non-objectivity: the property of causing bias due to allowing personal beliefs, judgments and emotions to transpire in language.

  Markers: words that express thoughts, beliefs, speculations, such as *think*, *hope*, *question*, etc.

- Uncertainty: indicating a questioning state of mind or an event taking place in the future for which sufficient data is not available to make a determination. Based on this definition, uncertainty is not always a marker of subjectivity. Often times, not inserting a hint of uncertainty may cause subjectivity, as in "Obama will win the election."

  Markers: adverbs such as *probably*, *apparently*, *perhaps*, etc

- Vagueness (present at the conceptual level): a concept that “does not have a precise definition” due to lack of a “well-defined frame of reference.”

  Markers: gradable words such as *little*, *popular*
- Ambiguity (present at the linguistic expression level): a concept that may have a different meaning based on the surrounding context.

  Markers: overloaded words that may have both objective and subjective meanings

Sentiment and subjectivity analysis can be performed at several levels. At the sense level, we can explore whether the sense of a given word is predominantly objective or subjective, positive, negative or neutral. At the next level, we can detect whether a word, or a multi-word expression encodes a private state. These micro-decisions then contribute to establishing at a macro-level understanding, whether a sentence, paragraph, or document exhibit subjectivity and / or polarity. From a natural language processing perspective, this process need not be incremental, from a limited view to an expanded one, but can also be reversed, from sentence and document level to extract proof to back up sentiment or subjectivity candidates at the expression level [79].

2.2. Challenges

- Lack of NLP infrastructure in most of the languages
- Lack of sentiment and subjectivity resources for most of the languages and high costs involved in generating them in isolation
- High variability in the types of text that can contain subjective language: from news to children stories (more complex, elevated to limited vocabulary, simplified), from blogs to tweets or rants (from proper formatted to free form, grammatical mistakes, elisions, short message language form)
- Isolating-synthetic-polysynthetic trait of languages (morpheme per word ratio): very isolating (Mandarin), rather isolating (English), rather synthetic (Japanese), etc. - making it impossible to have one fits all solutions
- Easier to detect subjectivity in larger spans of text versus short ones
- Subjective ambiguity: the property of a word to be subjective or objective given its context
- Implied subjectivity: expressed through word topology, irony, humor, etc

2.3. Applications

To date, a large number of text processing applications have used techniques for automatic sentiment and subjectivity analysis, including automatic expressive text-to-speech synthesis [**?**], tracking sentiment timelines in on-line forums and news [4, 41, 11], and mining opinions from product reviews [25, 51].

In many natural language processing tasks, subjectivity and sentiment classification has been used as a first phase filtering to generate more viable data. Research that benefited from this additional layering ranges from information retrieval, information extraction, question answering, to conversation summarization [13], and text semantic analysis [76, 18].

Information extraction. Information extraction (IE) aims to derive structured information from natural text. For systems of this type, text containing expressions of private states poses a difficult challenge, because words are most often used in their connoted sense (implying broader associations, thus nonfactual), compared to the denoted sense (i.e. the literal dictionary sense), which carries useful IE information. To exemplify, let us consider a system whose aim is to extract entities from the following text:

> A hiker in western Massachusetts was bitten by a venomous snake Saturday
> evening, then caught the snake and brought it to authorities. (2)
> Does he have the magic touch, or is he just a snake in the grass?

In the first sentence, the word "snake" is used in the denoted sense of *limbless scaled reptile (suborder Serpentes syn. Ophidia) with a long tapering body and with salivary glands often modified to produce venom which is injected through grooved or tubular fangs*[2], while in the second one the word connotes *a worthless or treacherous fellow*. An information extraction system should be able to discard nonfactual excerpts and focus on processing only the text that has the potential of providing valid data.

Numerous research papers [53, 55, 78] have shown that applying subjectivity analysis as a layering prior to the information extraction step allows the IE system to attain better precision.

Question answering. Question answering systems are used to provide an answer to a natural language question. They should be able to distinguish between a user's desire to obtain factual or

[2] `http://www.merriam-webster.com/dictionary/snake`

speculative answers, and thus direct the query to the pertinent subsystem. While a question such as *When did the first World War start?* is a close ended question, as it requires a factual finite answer, which can be automatically learned from encyclopedic knowledge sources, an open ended question such as *What is the meaning of life?* is more difficult to answer. A system for *automatic subjective question answering* may search in repositories of hand written answers to tackle such questions (from community question answering sites) or provide a collation or summarization of answers from different perspectives and / or various Internet sources.

Departing from the traditional factual question answering, the TREC opinion track [16] started to address the additional difficulties posed by automatically answering subjective questions under controlled settings. This trend has been picked up by additional researchers such as [38, 88], who are seeking to automatically detect whether the question posed by a user is subjective in nature, or [57] who use subjectivity analysis to separate opinion answers to multi-perspective questions.

Automatic summarization. Automatic summarization is the task of processing a text, gathering its most important aspects and presenting them in a succinct manner with the aid of a computer program. Ultimately, the desired outcome is a human readable summary. Often times, what is important in a text is not the dates, times, and other factual information, but rather the private states expressed by characters in novels or the points of tension in an email thread. Furthermore, [48, 19] have emphasized that subjective sentences attract more attention than those purely objective. For these reasons, an automatic summarization system may stand to gain from including subjective considerations in deciding the sentence set to be presented to the user, as [13] have shown.

Word sense disambiguation. WSD is concerned with detecting the correct sense held by a word in a sentence or span of text, when the word is ambiguous. Ultimately each correct decision in disambiguating a word in context contributes to establishing a coherent semantic meaning, thus allowing for the success of more abstract NLP tasks such as semantic similarity and relatedness. [76] have shown that subjectivity is a characteristic that is better defined at the sense level, as an ambiguous word may engulf both objective and subjective senses. Since one major problem for the task of WSD is the granularity of the reference lexicon, using an additional subjectivity dimension with only two states may allow for an easier path to identifying the correct sense as [1] have shown.

2.4. Metrics

For classification tasks, the evaluation metrics compare the results produced by a system against a gold standard, or a ground truth. Four terms are used: true negative (tn), false positive (fp), false negative (fn), true positive (tp) (see Table 2.2). Positive and negative refer to the prediction of the classifier, while true and false to the expectation of what the output should have been (gold standard).

Below I define the main metrics used in the proposed experiments:

- Precision. Measures the ability of a classifier to provide *only* relevant items:

  $Precision = \frac{tp}{tp+fp}$.
- Recall. Measures the ability of a classifier to provide *all* relevant items:

  $Recall = \frac{tp}{tp+fn}$.
- F-measure. It represents the weighted harmonic mean of precision and recall:

  $F = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$.
- Accuracy. It represents how well a classifier correctly identifies a condition (whether presence or absence):

  $Acc = \frac{tp+tn}{tp+tn+fp+fn}$.

	Predicted negative	Predicted positive
Actual negative	true negative (tn)	false positive (fp)
Actual positive	false negative (fn)	true positive (tp)

TABLE 2.2. Contingency table; actual is the true class or the *expectation*, while predicted is the decision of the classifier (*observation*).

CHAPTER 3

RELATED WORK

Before describing the work that has been carried out for multilingual sentiment and subjectivity analysis, I first briefly overview the main lines of research carried out on English, along with the most frequently used resources that have been developed for this language. Several of these English resources and tools have been used as a starting point to build resources in other languages, via cross-lingual projections or monolingual and multi-lingual bootstrapping. As described in more detail below, in cross-lingual projection, annotated data in a second language is created by projecting the annotations from a source (usually major) language across a parallel text. In multi-lingual bootstrapping, in addition to the annotations obtained via cross-lingual projections, mono-lingual corpora in the source and target languages are also used in conjunction with bootstrapping techniques such as co-training, which often lead to additional improvements.

While this chapter covers both sentiment and subjectivity analysis methods, the sentiment subfield is not explored exhaustively, but rather in conjunction with the subjectivity detection task. For this reason, methods that focus on sentiment or opinion (as they pertain to product reviews) are only included if they could be easily extrapolated to be employed for subjectivity classification in other languages as well. Since most languages other than a select hand few lack a strong natural language processing infrastructure, most methods developed for English cannot at the present time be implemented for other languages.

3.1. Sentiment and Subjectivity Analysis on English

3.1.1. Lexicons

One of the most frequently used lexicons is perhaps the subjectivity and sentiment lexicon provided with the OpinionFinder distribution [77]. The lexicon was compiled from manually developed resources augmented with entries learned from corpora. It contains 6,856 unique entries, out of which 990 are multi-word expressions. The entries in the lexicon have been labeled for part of speech as well as for reliability – those that appear most often in subjective contexts are *strong* clues of subjectivity, while those that appear less often, but still more often than expected by chance, are labeled *weak*. Each entry is also associated with a polarity label, indicating whether the corresponding word or phrase is positive, negative, or neutral. To illustrate, consider the following entry from the OpinionFinder lexicon *type=strongsubj word1=agree pos1=verb mpqapolarity=weakpos*, which indicates that the word *agree* when used as a *verb* is a strong clue of subjectivity and has a polarity that is weakly positive.

Another lexicon that has been often used in polarity analysis is the General Inquirer [56]. It is a dictionary of about 10,000 words grouped into about 180 categories, which have been widely used for content analysis. It includes semantic classes (e.g., animate, human), verb classes (e.g., negatives, becoming verbs), cognitive orientation classes (e.g., causal, knowing, perception), and other. Two of the largest categories in the General Inquirer are the valence classes, which form a lexicon of 1,915 positive words and 2,291 negative words.

SentiWordNet [19] is a resource for opinion mining built on top of WordNet, which assigns each synset in WordNet with a score triplet (positive, negative, and objective), indicating the strength of each of these three properties for the words in the synset. The SentiWordNet annotations were automatically generated, starting with a set of manually labeled synsets. Currently, SentiWordNet includes an automatic annotation for all the synsets in WordNet, totaling more than 100,000 words.

3.1.2. Corpora

Subjectivity and sentiment annotated corpora are useful not only as a means to train automatic classifiers, but also as resources to extract opinion mining lexicons. For instance, a large number of the entries in the OpinionFinder lexicon mentioned in the previous section were derived based on a large opinion-annotated corpus.

The MPQA corpus [80] was collected and annotated as part of a 2002 workshop on multiperspective question answering (thus the MPQA acronym). It is a collection of 535 English-language news articles from a variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.). The corpus was originally annotated at clause and phrase level, but sentence-level annotations associated with the dataset can also be derived via simple heuristics [80].

Another manually annotated corpus is the collection of newspaper headlines created and used during the recent SEMEVAL task on “Affective Text” [58]. The data set consists of 1000 test headlines and 200 development headlines, each of them annotated with the six Eckman emotions (anger, disgust, fear, joy, sadness, surprise) and their polarity orientation (positive, negative).

Two other data sets, both of them covering the domain of movie reviews, are a polarity data set consisting of 1,000 positive and 1,000 negative reviews, and a subjectivity data set consisting of 5,000 subjective and 5,000 objective sentences. Both data sets have been introduced in [48], and have been used to train opinion mining classifiers. Given the domain-specificity of these collections, they were found to lead to accurate classifiers for data belonging to the same or similar domains.

3.1.3. Tools

There are a large number of approaches that have been developed to date for sentiment and subjectivity analysis in English. The methods can be roughly classified into two categories: (1) rule-based systems, relying on manually or semi-automatically constructed lexicons; and (2) machine learning classifiers, trained on opinion-annotated corpora.

Among the rule-based systems, one of the most frequently used is OpinionFinder [77], which automatically annotates the subjectivity of new text based on the presence (or absence) of words or phrases in a large lexicon. Briefly, the OpinionFinder high-precision classifier relies on three main heuristics to label subjective and objective sentences: (1) if two or more strong subjective expressions occur in the same sentence, the sentence is labeled *Subjective*; (2) if no strong subjective expressions occur in a sentence, and at most two weak subjective expressions occur in the previous, current, and next sentence combined, then the sentence is labeled *Objective*; (3) otherwise, if none of the previous rules apply, the sentence is labeled *Unknown*. The classifier uses the clues from a subjectivity lexicon and the rules mentioned above to harvest subjective and objective sentences from a large amount of unannotated text; this data is then used to automatically identify a set of extraction patterns, which are then used iteratively to identify a larger set of subjective and objective sentences.

In addition to the high-precision classifier, OpinionFinder also includes a high-coverage classifier. This high-precision classifier is used to automatically produce an English labeled data set, which can then be used to train a high-coverage subjectivity classifier.

When evaluated on the MPQA corpus, as reported by [77], the high-precision classifier was found to lead to a precision of 86.7% and a recall of 32.6%, whereas the high-coverage classifier has a precision of 79.4% and a recall of 70.6% (see Table 3.1).

	P	R	F
high-precision	86.7%	32.6%	47.4%
high-coverage	79.4%	70.6%	74.7%

TABLE 3.1. Precision (P), Recall (R) and F-measure (F) for the two OpinionFinder classifiers, as measured on the English MPQA corpus

Another unsupervised system worth mentioning, this time based on automatically labeled words or phrases, is the one proposed in [71], which builds upon earlier work by [24]. Starting with two reference words, "excellent" and "poor," Turney classifies the polarity of a word or phrase by measuring the fraction between its pointwise mutual information (PMI) with the positive reference ("excellent") and the PMI with the negative reference ("poor").[1] The polarity scores assigned in this way are used to automatically annotate the polarity of product, company, or movie reviews. Note that this system is completely unsupervised, and thus particularly appealing for application to other languages.

Finally, when annotated corpora is available, machine-learning methods are a natural choice for building subjectivity and sentiment classifiers. For example, Wiebe at al. [75] used a data set manually annotated for subjectivity to train a machine learning classifier, which led to significant improvements over the baseline. Similarly, starting with semi-automatically constructed data sets,

[1] The PMI of two words w_1 and w_2 is defined as the probability of seeing the two words together divided by the probability of seeing each individual word: $PMI(w_1, w_2) = \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$

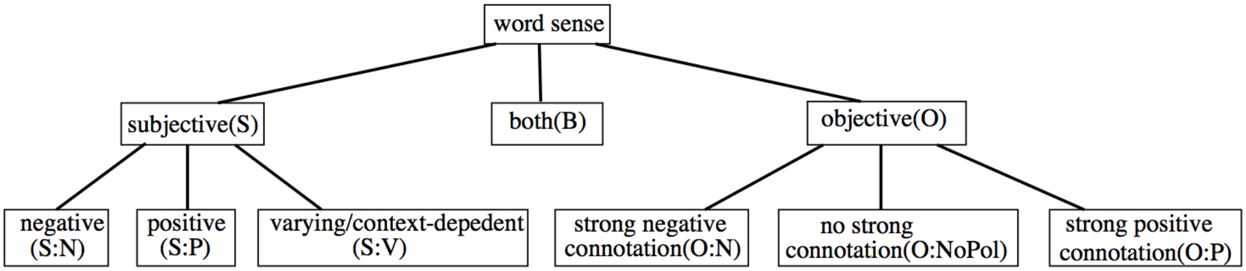

word sense
subjective(S)
both(B)
objective(O)
negative (S:N)
positive (S:P)
varying/context-depedent (S:V)
strong negative connotation(O:N)
no strong connotation(O:NoPol)
strong positive connotation(O:P)

FIGURE 3.1. Sense level hierarchical classification proposed by [60]

Pang and Lee [48] built classifiers for subjectivity annotation at sentence level, as well as a classifier for sentiment annotation at document level. To the extent that annotated data is available, such machine-learning classifiers can be used equally well in other languages.

3.2. Sense-level Annotations

Sense level subjectivity analysis is a task that was first proposed and explored by [76] in 2006. The authors started from two questions, namely whether senses can be annotated for subjectivity, and whether this additional information can prove beneficial in the task of word sense disambiguation. Upon conducting a manual annotation study, they showed that senses can be labeled for subjectivity with a good agreement (85.5%, κ = 0.74), and when removing uncertain cases, the agreement increases to 95% (κ = 0.9). Motivated by these results, the authors propose an automatic way of detecting the subjectivity of senses by pairing the similarity between a word sense and words that are distributionally similar to it with the fact that they appear in a manually annotated subjective expression in the MPQA corpus [80]. When considering pairing distributionally similar words with only the sense with which they achieve maximum similarity, a precision and recall level of 0.5 is reached. This automatic sense level subjectivity system is used to append its prediction to a vectorial space on which a word sense disambiguation engine is trained. Upon considering this additional information, the system obtains an error rate reduction of 4.3% for subjectivity ambiguous words, and 2.2 % error rate reduction for the words with no subjective senses; a small degradation or no change is observed for individual entries in the latter group.

Motivated by the previous study, Su and Markert [60] also introduce an annotation scheme for subjectivity and polarity at the word sense level. The approach proposes a hierarchical classification, first for subjectivity (using the tagset *subjective*, *objective* and *both*), and then for polarity (see Figure 3.1). Unlike previous subjectivity research (such as [24, 65]), Su and Markert [60] do not only consider the polarity of subjective senses, but rather assume that objective senses may also encode polarity based on their connotated meaning in the Western culture. As an example, they list the following objective sense extracted from WordNet:

> **war, warfare** - the waging for armed conflict against an enemy; "thousands of people were killed in the war"

They label this sense as negative, as "war" is a negative event, based on their definition of connotated polarity "senses that do not describe or express an emotion or judgment but whose presence in a text would give it a negative flavor."

However, upon conducting annotation studies on the Micro-WNOp corpus[2], the only reliable interannotator agreement (higher than 0.8%) is observed for the traditional categories, namely subjective further classified as positive or negative, and objective (with no polarity). When removing the hierarchical approach, and annotating for subjectivity or for polarity alone, the agreement increases to 90.1% (κ = 0.79) for subjectivity and 89.1% (κ = 0.83) for polarity. This seems to indicate that deciding on the polarity of *objective* senses is not a well defined task, as connotations engage in text based on the surrounding context. When only the definition and the synonyms / antonyms of a word are available, connotations are dependent on the mood and the neural network of the annotator (its synapses and brain activation patterns), making the annotations unreliable. As such, "war" may not participate in eliciting a subjective or negative feeling at all, for example:

Before the dawn of civilization, war likely consisted of small-scale raiding.[3] (3)

Encouraged by the high agreement results for subjectivity classification at the sense level, [61] introduce both supervised and unsupervised methods tested on the Micro-WNOp corpus proposed earlier. Using 10 fold cross validation on the unigram feature space extracted from the

[2] Available at `http://www.unipv.it/wnop/micrownop.tgz`

glosses available in the Micro-WNOp corpus paired with Naïve Bayes learners, they are able to obtain an accuracy of 74.6% (to compare with a majority class baseline of 66.3%). Upon also adding part-of-speech information and WordNet relations (such as antonym, similar-to, derived-from, attribute, etc.) the accuracy improves by 2 percentage points. When casting the task of *sense* level subjectivity detection as *sentence* level classification (where a sentence is approximated with a sense's gloss), by constructing a unigram training space from sentences annotated for subjectivity from the MPQA corpus [80] or the Movie dataset [48], and learning Naïve Bayes classifiers, the results fall below the baseline. Comparable results to cross-validation are obtained when using the subjectivity lexicon included with the OpinionFinder distribution [77] in either a supervised or rule-based approach. For the supervised approach, the label of a gloss is given based on the cumulative score of words appearing in the subjectivity lexicon; *weak* entries participate with a score of 1, while *strong* ones with a score of 2; the system is then trained on the labeled samples. For the rule-based approach, the cumulative score alone dictates the sense classification. All methods that use the OpinionFinder subjectivity lexicon achieve an accuracy above 74%, yet the machine learning experiments performs better (75.7% when only unigrams are considered, and 76.9% when also adding part-of-speech and WordNet relations information). The General Inquirer (GI) lexicon [56] provides an accuracy below 70% in both supervised and unsupervised experiments. This implies that for the task of sense level subjectivity classification, the OpinionFinder subjectivity lexicon is better suited, as it focuses on *subjectivity* instead of *polarity*. Also, using unsupervised learning methods can reach within approximately 2% accuracy of the supervised ones.

Esuli and Sebastiani [19] automatically label the entire WordNet with subjectivity / polarity information by starting out with a manually selected small set of words annotated for polarity. The authors use 3 training pools: positive, negative and objective. These pools are created as a result of a bootstrapping process starting with the same seeds from [71] and augmented based on relationships they have with other WordNet synsets (such as antonymy and see also). For a given synset, a vector is created from the gloss (using cosine normalized $tf.idf$[4]). These vectors help train a series of 8 classifiers that perform majority voting when new unlabeled synset samples are

[4] Term frequency - inverse document frequency, a common statistical measure that is used to identify the importance of a word in a document or copus.

provided. The authors do not present any evaluations against a gold standard. In [20], they expand their method by conferring polarity labels based on scores provided by a PageRanking algorithm applied to the WordNet graph and compare the scores obtained upon convergence with those from the Micro-WNOp corpus.

However, all these research approaches were carried out in English. Turning now towards subjectivity research at the sense level in other languages, [21] did consider transferring automatically inferred polarity annotations at the sense level from the English SentiWordNet to Italian using MultiWordNet [49][5] for the purpose of opinion extraction. Since their task was to annotate expressions of opinion, such as the *opinion* itself, the *holder* (the person expressing an opinion) and the *target* (the entity the opinion is about), they flattened the sense-level annotations by devising a cumulative *word*-level score representing the summation over all the senses of the positive and negative scores.

Su and Markert [63] used sense-level subjectivity to tackle the cross-lingual lexical substitution task, namely to provide the correct translation for a target word that appears in numerous contexts under different senses; in their experiments the contexts are expressed in English, while the target word is translated into Chinese). They show that adding manual or automatic subjectivity annotations to subjectivity-ambiguous words and considering the annotations as features in a machine learning setup improves accuracy by 9.4% when using manual tagging and 2.2% when using automatic labeling. While their work is not focused on sentiment or subjectivity detection, it is based on the intuition that the subjectivity label of senses tends to remain constant across language boundaries; however, they do not carry out experiments to verify that this is the case.

3.3. Word and Phrase-level Annotations

The development of resources and tools for sentiment and subjectivity analysis often starts with the construction of a lexicon, consisting of words and phrases annotated for sentiment or subjectivity. Such lexicons are successfully used to build rule-based classifiers for automatic opinion annotation, by primarily considering the presence (or absence) of the lexicon entries in a text.

[5] MultiWordNet is a lexical resource that is part of the multilingual WordNet family, and thus follows the WordNet structure and alignment and is developed for Italian.

There are three main directions that have been considered so far for word and phrase level annotations: (1) manual annotations, which involve human judgment of selected words and phrases; (2) automatic annotations based on knowledge sources such as dictionaries; and (3) automatic annotations based on information derived from corpora.

3.3.1. Dictionary-based

Using a translation technique, Kim and Hovy [34] build a lexicon for German starting with a lexicon in English, this time focusing on polarity rather than subjectivity. They use an English polarity lexicon semi-automatically generated starting with a few seeds and using the WordNet structure [45]. Briefly, for a given seed word, its synsets and synonyms are extracted from WordNet, and then the probability of the word belonging to one of the three classes is calculated based on the number and frequency of seeds from a particular class appearing within the word's expansion. This metric thus represents the closeness of a word to the seeds. Using this method, Kim and Hovy are able to generate an English lexicon of about 1,600 verbs and 3,600 adjectives, classified as positive or negative based on their polarity.

The lexicon is then translated into German, by using an automatically generated translation dictionary obtained from the European Parliament corpus via word alignment [47]. To evaluate the quality of the German polarity lexicon, the entries in the lexicon were used in a rule-based system that was applied to the annotation of polarity for 70 German emails. Overall, the system obtained an F-measure of 60% for the annotation of positive polarity, and 50% for the annotation of negative polarity.

A similar bootstrapping technique was used by Pitel and Grefenstette [50], for the construction of affective lexicons for French. They classify words into 44 affect classes (e.g., *morality*, *love*, *crime*, *insecurity*), each class being in turn associated with a positive or negative orientation. Starting with a few seed words (two to four seed words for each affective dimension), they use synonym expansion to automatically add new candidate words to each affective class. The new candidates are then filtered based on a measure of similarity calculated with latent semantic analysis, and machine learning trained on seed data. Using this method, Pitel and Grefenstette are able to generate a French affective lexicon of 3,500 words, which is evaluated against a gold standard data set consisting of manually annotated entries. As more training samples are available in the training lexicon, the F-measure classification increases from 12% to 17%, up to a maximum of 27% F-measure for a given class.

3.3.2. Corpus-based

In addition to dictionaries, textual corpora were also found useful to derive subjectivity and polarity information associated with words and phrases. Much of the corpus-based research carried out to date follows the work of Turney [71] (see Section 3.1.3), who presented a method to measure the polarity of a word based on its PMI association with a positive or a negative seed (e.g., *excellent* and *poor*).

In [29], Kaji and Kitsuregawa propose a method to build sentiment lexicons for Japanese, by measuring the strength of association with positive and negative data automatically collected from Web pages. First, using structural information from the layout of HTML pages (e.g., list markers or tables that explicitly indicate the presence of the evaluation sections of a review, such as "pros"/"cons", "minus"/"plus", etc.), as well as Japanese-specific language structure (e.g., particles used as topic markers), a corpus of positive and negative statements is automatically mined from the Web. Starting with one billion HTML documents, about 500,000 polar sentences are collected, with 220,000 being positive and the rest negative. Manual verification of 500 sentences, carried out by two human judges, indicated an average precision of 92%, which shows that reasonable quality can be achieved using this corpus construction method.

Next, Kaji and Kitsuregawa use this corpus to automatically acquire a set of polar phrases. Starting with all the adjectives and adjectival phrases as candidates, they measure the chi-squared and the PMI between these candidates and the positive and negative data, followed by a selection of those words and phrases that exceed a certain threshold. Through experiments, the PMI measure was found to work better as compared to chi-squared. The polarity value of a word or phrase based on PMI is defined as:

$$PV_{PMI}(W) = PMI(W, pos) - PMI(W, neg)$$

where

$$PMI(W,pos) = log_2 \frac{P(W,pos)}{P(W)P(pos)} \qquad PMI(W,neg) = log_2 \frac{P(W,neg)}{P(W)P(neg)}$$

pos and neg representing the positive and negative sentences automatically collected from the Web.

Using a data set of 405 adjective phrases, consisting of 158 positive phrase, 150 negative, and 97 neutral, Kaji and Kitsuregawa are able to build a lexicon ranging from 8,166 to 9,670 entries, depending on the value of the threshold used for the candidate selection. The precision for the positive phrases was 76.4% (recall 92.4%) when a threshold of 0 is used, and went up to 92.0% (recall 65.8%) when the threshold is raised to 3.0. For the same threshold values, the negative phrases had a precision ranging from 68.5% (recall 84.0%) to 87.9% (recall 62.7%).

Another corpus-based method for the construction of polarity lexicons in Japanese, this time focusing on domain-specific propositions, is proposed in [30]. Kanayama and Nasukawa introduce a novel method for performing domain-dependent unsupervised sentiment analysis through the automatic acquisition of polar atoms in a given domain by building upon a domain-independent lexicon. In their work, a polar atom is defined as "the minimum human-understandable syntactic structures that specify the polarity of clauses," and it typically represents a tuple of polarity and a verb or an adjective along with its optional arguments. The system uses both intra- and inter-sentential coherence as a way to identify polarity shifts, and automatically bootstraps a domain-specific polarity lexicon.

First, candidate propositions are identified by using the output of a full parser. Next, sentiment assignment is performed in two stages. Starting from a lexicon of pre-existing polar atoms based on an English sentiment lexicon, the method finds occurrences of the entries in the propositions extracted earlier. These propositions are classified as either positive or negative based on the label of the atom they contain, or its opposite in case a negation is encountered. The next step involves the extension of the initial sentiment labeling to those propositions that are not labeled. To this end, context coherency is used, which assumes that in a given context the polarity will not shift unless an adversative conjunction is encountered, either between sentences and/or within sentences. Finally, the confidence of each new polar atom is calculated, based on its total number of occurrences in positive and negative contexts.

The method was evaluated on Japanese product reviews extracted from four domains: digital cameras, movies, mobile phones and cars. The number of reviews in each corpus ranged from 155,130 (mobile phones) to 263,934 (digital cameras). Starting with these data sets, the method is able to extract 200-700 polar atoms per domain, with a precision evaluated by human judges ranging from 54% for the mobile phones corpus to 75% for the movies corpus.

Kanayama and Nasukawa's method is similar to some extent to an approach proposed earlier by Kobayashi et al., which extracts opinion triplets from Japanese product reviews mined from the Web [35]. An opinion triplet consists of the following fields: product, attribute and value. The process involves a bootstrapping process consisting of two steps. The first step consists of the generation of candidates based on a set of co-occurrence patterns, which are applied to a collection of Web reviews. Three dictionaries that are updated at the end of each bootstrapping iteration are also provided (dictionaries of subjects, attributes, and values). Once a ranked list of candidates is generated, a human judge is presented with the top ranked candidates for annotation. The manual step involves identifying the attributes and their values and updating their corresponding dictionaries with the newly extracted entities.

For the experiments, Kobayashi et al. use two data sets, consisting of 15,000 car reviews and 10,000 game reviews respectively. The bootstrapping process starts with a subject dictionary of 389 car names and 660 computer games names, an initial attribute list with seven generic descriptors (e.g., cost, price, performance), and a value list with 247 entries (e.g., good, beautiful, high). Each extraction pattern is scored based on the frequency of the extracted expressions and their reliability. For the evaluation, a human annotator tagged 105 car reviews and 280 computer game reviews, and identified the attributes and their corresponding values. Overall, using the semi-automatic system, Kobayashi et al. found that lexicons of opinion triplets can be built eight times faster as compared to a fully manual set-up. Moreover, the semi-automatic system is able to achieve a coverage of 35-45% with respect to the manually extracted expressions, which represents a significant coverage.

The semantic orientation of phrases in Japanese is also the goal of the work of [66] and [64], both using an expectation maximization model trained on annotated data. Takamura et al.

consider the task of finding the polarity of phrases such as "light laptop," which cannot be directly obtained from the polarity of individual words (since, in this case, both "light" and "laptop" are neutral). On a data set of 12,000 adjective-noun phrases drawn from a Japanese newspaper, they found that a model based on triangle and "U-shaped" graphical dependencies leads to an accuracy of approximately 81%.

Suzuki et al. target instead evaluative expressions, similar to those addressed by [35]. They use an expectation maximization algorithm and a Naïve Bayes classifier to bootstrap a system to annotate the polarity of evaluative expressions consisting of subjects, attributes and values. Using a data set of 1,061 labeled examples and 34,704 unlabeled examples, they obtain an accuracy of 77%, which represents a significant improvement over the baseline of 47% obtained by assigning the majority class from the set of 1,061 labeled examples.

Finally, another line of work concerned with the polarity analysis of words and phrases is presented in [11]. Instead of targeting the derivation of subjectivity or sentiment lexicon in a new language, the goal of Bautin et al.'s work is to measure the polarity of given entities (e.g., George Bush, Vladimir Putin) in a text written in a target language. Their approach relies on the translation of documents (e.g., newswire, European parliament documents) from the given language into English, followed by a calculation of the polarity of the target entity by using association measures between the occurrence of the entity and positive/negative words from a sentiment lexicon in English.

The experiments presented in [11] focus on nine different languages (Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Spanish), and fourteen entities covering country and city names. They show that large variations can be achieved in the measures of polarity or subjectivity of an entity across languages, ranging from very weak correlations (close to 0), to strong correlations (0.60 and higher). For instance, an aggregation of all the polarity scores measured for all fourteen entities in different languages leads to a low correlation of 0.08 between mentions of such entities in Japanese and Chinese text, but as high as 0.63 when the mentions are collected from French and Korean texts.

3.4. Sentence-level Annotations

Corpus annotations are often required either as an end goal for various text processing applications (e.g., mining opinions from the Web; classification of reviews into positive and negative; etc.), or as an intermediate step toward building automatic subjectivity and sentiment classifiers. Work in this area has considered annotations at either sentence or document level, depending mainly on the requirements of the end application (or classifier). The annotation process is typically done following one of three methods: (1) dictionary-based, consisting of rule-based classifiers relying on lexicons built with one of the methods described in the previous section; (2) corpus-based, consisting of machine learning classifiers trained on pre-existing annotated data; or (3) hybrid, combining elements from both the dictionary and the corpus-based methods.

3.4.1. Dictionary-based

Rule-based classifiers, such as the one introduced Riloff and Wiebe in [52], can be used in conjunction with any opinion lexicon to develop a sentence-based classifier. These classifiers mainly look for the presence (or absence) of lexicon clues in the text, and correspondingly decide on the classification of a sentence as subjective/objective or positive/negative.

A lexicon approach is used for the classification of polarity for sentences in Japanese [31]. Kanayama et al. use a machine translation system based on deep parsing to extract "sentiment units" with high precision from Japanese product reviews, where a sentiment unit is defined as a touple between a sentiment label (positive or negative) and a predicate (verb or adjective) with its argument (noun). The sentiment analysis system uses the structure of a transfer-based machine translation engine, where the production rules and the bilingual dictionary are replaced by sentiment patterns and a sentiment lexicon, respectively.

The system is ultimately able to not only mine product reviews for positive/negative product attributes, but also to provide a user friendly interface to browse product reviews. The sentiment units derived for Japanese are used to classify the polarity of a sentence, using the information drawn from a full syntactic parser in the target language. Using about 4,000 sentiment units, when evaluated on 200 sentences, the sentiment annotation system was found to have high precision (89%) at the cost of low recall (44%).

3.4.2. Corpus-based

Once a corpus annotated at sentence level is available, with either subjectivity or polarity labels, a classifier can be easily trained to automatically annotate additional sentences.

This is the approach taken by Kaji and Kitsuregawa [28, 29], who collect a large corpus of sentiment-annotated sentences from the Web, and subsequently use this data set to train sentence-level classifiers. Using the method described in Section 3.3.2, which relies on structural information from the layout of HTML pages, as well as Japanese-specific language structure, Kaji and Kitsuregawa collect a corpus of approximately 500,000 positive and negative sentences from the Web. The quality of the annotations was estimated by two human judges, who found an average precision of 92% as measured on a randomly selected sample of 500 sentences.

A subset of this corpus, consisting of 126,000 sentences, is used to build a Naïve Bayes classifier. Using three domain specific data sets (computers, restaurants and cars), automatically collected by selecting manually annotated reviews consisting of only one sentence, the precision of the classifier was found to have an accuracy ranging between 83% (computers) and 85% (restaurants), which is comparable to the accuracy obtained by training on in-domain data. These results demonstrate the quality of the automatically built corpus, and the fact that it can be used to train reliable sentence-level classifiers with good portability to new domains.

3.4.3. Hybrid

Mukund and Srihari [46] propose a monolingual hybrid model for subjectivity detection in Urdu at the sentence level. They extract 500 articles from BBC Urdu written after 2003, based on a set of sentiment keywords such as anger, love, etc. Upon manual annotation of this set using the MPQA annotation framework [80], approximately 700 sentences are subjective and the remaining 6000 are objective. These sets are split into 70% training (used by the co-training step) and 30% testing. In order to generate positive and negative class data suitable for learning using a supervised machine learning algorithm, they introduce a co-training step (inspired by [12]) that builds on the knowledge extracted from the positive class (composed of subjective sentences) and its dissimilarity with the unlabeled data (see Figure 3.2). The features employed in their model are: word unigrams, part-of-speech, words appearing in the WordNet Affect emotion list [59], 23 syntactic patterns and confidence words. The latter are words weighted based on a Wilson Proportion Estimate metric [81] which computes a level of confidence between a given word and the subjective / objective class. This is done by considering the number of documents containing the word associated with the two classes. Experiments are conducted on the newswire data using a vector space model (VSM) and a support vector machine (SVM) algorithm, and obtain an F-measure of 86% when using the former. While the results seem to indicate that a VSM model is more robust when dealing with a limited amount of annotated data in Urdu, the settings (i.e. feature space and number of training samples) are not equivalent between the two classifiers in order to fully support this assertion. They obtain an average F-measure of 78% on the IMDB dataset (introduced in Section 3.1.2).

3.5. Document-level Annotations

Natural language applications, such as review classification or Web opinion mining, often require corpus-level annotations of subjectivity and polarity. In addition to sentence-level annotations, described in the previous section, there are several methods that have been proposed for the annotation of entire documents. As before, the two main directions of work have considered: (1) dictionary-based annotations, which assume the availability of a lexicon, and (2) corpus-based annotations, which mainly rely on classifiers trained on labeled data.

3.5.1. Dictionary-based

Perhaps the simplest approach for document annotations is to use a rule-based system based on the clues available in a language-specific lexicon. One of the methods proposed by Wan [73] consists of annotating Chinese reviews by using a polarity lexicon, along with a set of negation words and intensifiers. The lexicon contains 3,700 positive terms, 3,100 negative words, and 148 intensifier terms, all of them collected from a Chinese vocabulary for sentiment analysis released by HowNet, as well as 13 negation terms collected from related research. Given this lexicon, the polarity of a document is annotated by combining the polarity of its constituent sentences, where in turn the polarity of a sentence is determined as a summation of the polarity of the words found

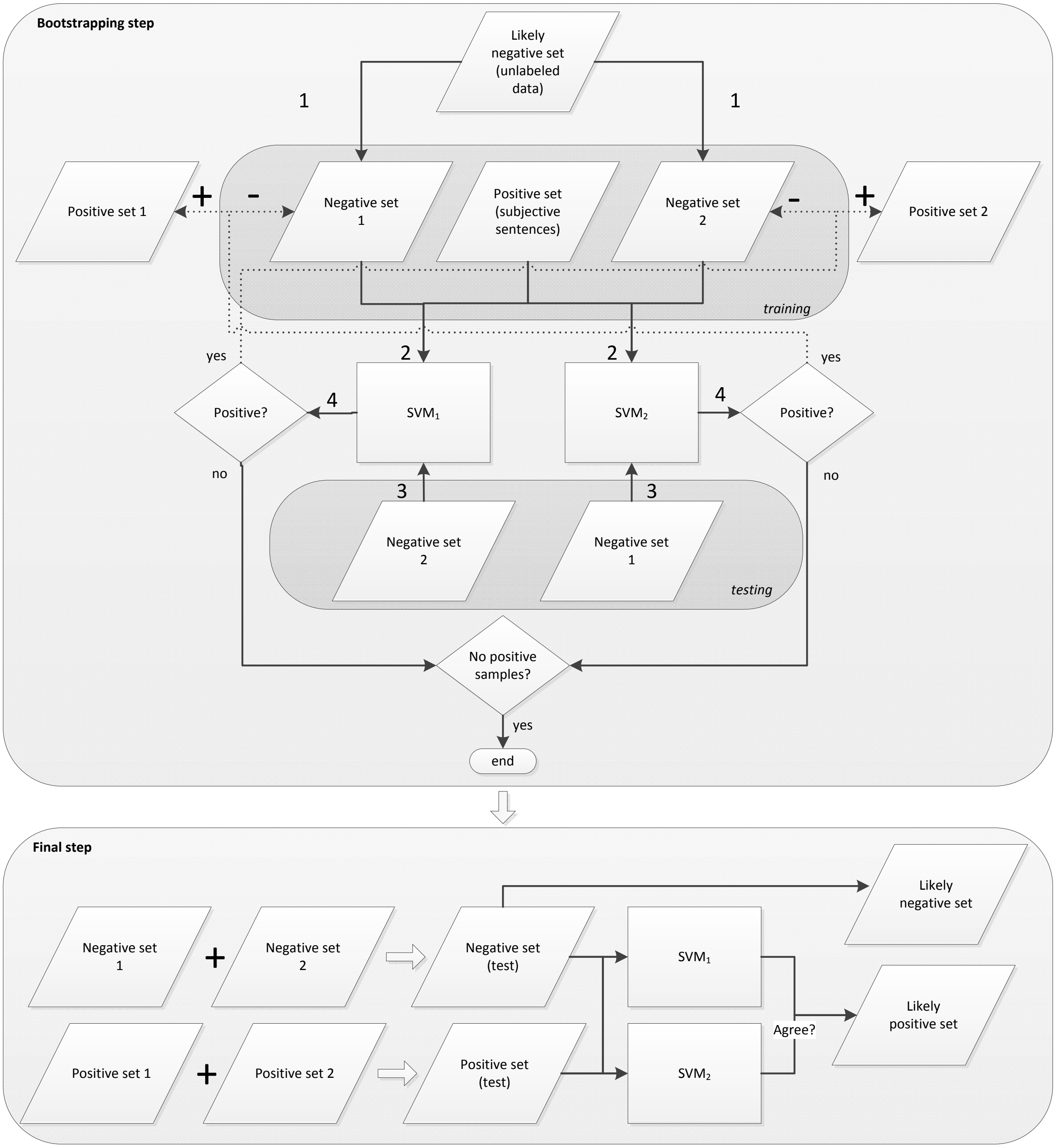


FIGURE 3.2. Sentence level monolingual co-training.

in the sentence. When evaluated on a data set of 886 Chinese reviews, this method was found to give an overall accuracy of 74.3%.

The other method proposed by Wan [73] is to use machine translation to translate the Chinese reviews into English, followed by the automatic annotation of the English reviews using a rule-based system relying on English lexicons. Several experiments are run with two commercial machine translation systems, using the OpinionFinder polarity lexicon (see Section 3.1.1). For the same test set mentioned before, the translation method achieves an accuracy of up to 81%, significantly higher than the one achieved by directly analyzing the reviews using a Chinese lexicon. Moreover, an ensemble combining different translations and methods leads to an even higher accuracy of 85%, demonstrating that a combination of different knowledge sources can exceed the performance obtained with individual resources.

Another approach, proposed by Zagibalov and Carroll [87], consists of a bootstrapping method to label the polarity of Chinese text by iteratively building a lexicon and labeling new text. The method starts by identifying "lexical items" in text, which are sequences of Chinese characters that occur between non-character symbols and which include a negation and an adverbial; a small hand-picked list of six negations and five adverbials is used, which increase the portability of the method to other languages. In order to be considered for candidacy in the seed list, the lexical item should appear at least twice in the data that is being considered.

Next, "zones" are identified in the text, where a zone is defined as the sequence of characters occurring between punctuation marks. The sentiment associated with an entire document is calculated as the difference between the number of positive and negative zones that the review entails. In turn, the sentiment of a zone is computed by summing the polarity scores of their component lexical items. Finally, the polarity of a lexical item is proportional with the square of its length (number of characters), and with is previous polarity score, while being inversely proportional to the length of the containing zone. This score is multiplied by -1 in case a negation precedes the lexical item.

The bootstrapping process consists of iterative steps that result in an incrementally larger set of seeds, and an incrementally larger number of annotated documents. Starting with a seed set consisting initially of only one adjective ("good"), new documents are annotated as positive and negative, followed by the identification of new lexical items occurring in these documents that can be added to the seed set. The addition to the seed set is determined based on the frequency of the lexical item, which has to be at least three time larger in the positive (negative) documents for it to be considered. The bootstrapping stops when over two runs no new seeds are found.

The method was evaluated over a balanced corpus of Chinese reviews compiled from ten different domains. The average accuracy at document level was measured at 83%. Moreover, the system was also able to extract a set of 50-60 seeds per domain, which may be helpful for other sentiment annotation algorithms.

Another method, used by Kim and Hovy [34], consists of the annotation of German documents using a lexicon translated from English. A lexicon construction method, described in detail in Section 3.3.1, is used to generate an English lexicon of about 5,000 entries. The lexicon is then translated into German, by using an automatically generated translation dictionary obtained from the European Parliament corpus using word alignment. The German lexicon is used in a rule-based system that is applied to the annotation of polarity for 70 German emails. Briefly, the polarity of a document is decided based on heuristics: a number of negative words above a particular threshold renders the document negative, whereas a majority of positive words triggers a positive classification. Overall, the system obtained an F-measure of 60% for the annotation of positive polarity, and 50% for the annotation of negative polarity.

3.5.2. Corpus-based

The most straight-forward approach for corpus-based document annotation is to train a machine learning classifier, assuming that a set of annotated data already exists. Li and Sun [39] use a data set of Chinese hotel reviews, on which they apply several classifiers, including SVM, Naïve Bayes and maximum entropy. Using a training set consisting of 6,000 positive reviews and 6,000 negative reviews and a test set of 2,000 positive reviews and 2,000 negative reviews, they obtain an accuracy of up to 92%, depending on the classifier and on the features used. These experiments demonstrate that if enough training data are available, it is relatively easy to build accurate sentiment classifiers.

A related, yet more sophisticated technique is proposed in [74], where a co-training approach is used to leverage resources from both a source and a target language. The technique is tested on the automatic sentiment classification of product reviews in Chinese. For a given product review in the target language (Chinese), an alternative view is obtained in another language (English) via machine translation. The algorithm then uses two SVM classifiers, one in Chinese and one in English, to start a co-training process that iteratively builds a sentiment classifier. Initially, the training data set consists of a set of labeled examples in Chinese and their English translations. Next, the first iteration of co-training is performed, and a set of unlabeled instances is classified and added to the training set if the labels assigned in the models built on the languages agree. The newly labeled instances are used to re-train the two classifiers at the next iteration. Reviews with conflicting labels are not considered. As expected, the performance initially grows with the number iterations, followed by a degradation when the number of erroneously labeled instances exceeds a certain threshold. The best results are reported at the 40th iteration, for an overall F-measure of 81%, after adding five negative and five positive reviews at each iteration. The method is successful because it makes use of both cross-language and within-language knowledge.

CHAPTER 4

SENSE-LEVEL MULTILINGUAL SUBJECTIVITY

Recent research on English word sense subjectivity has shown that the subjective aspect of an entity is a characteristic that is better delineated at the sense level, instead of the traditional word level. This chapter explores whether senses aligned across languages exhibit this trait consistently, and if this is the case, further investigates how this property can be leveraged in an automatic fashion. A manual annotation study is first conducted to gauge whether the subjectivity trait of a sense can be robustly transferred across language boundaries. Then, an automatic framework is introduced that is able to predict subjectivity labeling for unseen senses using either cross-lingual or multilingual training enhanced with bootstrapping. The experiments conducted suggest that the multilingual model consistently outperforms the cross-lingual one, with an accuracy of over 73% across all iterations.

I seek to answer the following questions. First, for word senses aligned across languages, is their subjectivity content consistent, or in other words, does a subjective sense in language A map to a subjective sense in language B (and similarly for an objective sense)? Second, can a multilingual framework that can automatically discover new subjective/objective senses starting with a limited amount of annotated data be employed? The answer to the first question is sought by conducting a manual annotation study in Section 4.1. For the second question, I propose two models (see Section 4.2), one cross-lingual and one multilingual, which are able to simultaneously use information extracted from several languages when making subjectivity sense-level predictions.

Parts of this research were previously published in [7, 9].

4.1. Sense Level Subjectivity Consistency Across Languages: Annotation Study

To answer the first question, a case study in subjectivity sense transfer across languages focusing on English and Romanian is framed.

Let us consider a sense-level aligned multilingual resource such as WordNet. WordNet [45] was first developed for English, and is a lexical resource that maintains semantic relationships between basic units of meaning, or *synsets*. A synset groups together senses of different words that share a very similar meaning. Due to its particular usefulness for NLP tasks, numerous independent non-commercial projects[1] have replicated its structure in over 50 languages, while maintaining alignment with the original WordNet and allowing for sense-level mapping across languages.

Our experiments use the English [45] and the Romanian [69] versions of WordNet, which contain 117659[2] and 58725[3] synsets, respectively. Many of these are aligned at the synset level.

In order to infuse subjectivity information into the model, a list of 128 English words[4] accounting for 580 senses (with an average polysemy of 4.6) is used, based on sense-level manually annotated subjectivity data from [76] and [1], as well as a list of 48 additional words (obtained through private communication with the aforementioned authors). Their equivalent into Romanian is obtained by traversing the WordNet structure. A native speaker of Romanian (who participated in previous subjectivity annotations studies) was asked to annotate the Romanian data, by being presented with the *gloss* (definition) and the *synset* of each given sense from the Romanian WordNet. The annotator agreement between the English and the Romanian subjectivity labels ranged from 84% (for the [1] dataset) to 90% (for the [76] dataset). When excluding senses that had both subjective and objective uses (labeled as *both*) in either of the languages, the annotator agreement becomes 87%, with Cohen's $\kappa = 0.74$ for the first dataset, and 94.7% with $\kappa = 0.88$ for the second one, indicating good to very good agreement. These findings support the hypothesis that the subjectivity of a sense maintains itself across language boundaries. Furthermore, they indicate that senses aligned across languages may represent vessels of subjectivity transfer into other languages, thus providing an anchor to generating subjectivity annotated lexica in a target language. Since not all senses have the same subjectivity label across languages, the following section explores in more detail the various scenarios encountered.

4.1.1. Differences between Languages

There were several examples where the subjectivity label changed between languages. For instance, the fourth sense of the noun *argument*, as listed in Table 4.1, is marked in the English

[1] `http://www.globalwordnet.org/gwa/wordnet_table.htm`
[2] `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`
[3] `http://www.racai.ro/wnbrowser/Help.aspx`
[4] The tag set employed by the annotators consists of *subjective*, *objective*, and *both*.

Differences between languages			
argument	Gloss	En: a summary of the subject or plot of a literary work or play or movie "the editor added the argument to the poem"	Ro: redare-prezentare pe scurt-scrisă sau orală- a ideilor unei lucrări- ale unei expuneri etc. ($translation$) short summary, oral or in writing, of the ideas presented in a literary work
	Synset	En: argument, literary argument	Ro: rezumat ($translation$) summary
decide	Gloss	En: influence or determine "The vote in New Hampshire often decides the outcome of the Presidential election"	Ro: a exercita o influenţă - a determina ($translation$) to exercise influence - to determine
	Synset	En: decide	Ro: influenţa; decide; hotărî ($translation$) influence; decide; determine
WordNet Granularity			
free	Gloss	En: able to act at will; not hampered; not under compulsion or restraint; "free enterprise"; "a free port"; "a free country"; "I have an hour free"; "free will"; "free of racism"; "feel free to stay as long as you wish"; "a free choice"	Ro: (Despre oameni) Care are posibilitatea de a acţiona după voinţa sa - de a face sau de a nu face ceva; ($translation$) (About people) Someone who can act according to his will - who can do or not do something
	Synset	En: free	Ro: liber ($translation$) free

TABLE 4.1. *Sources of conflict in cross-lingual subjectivity transfer.* Definitions and synonyms of the fourth sense of the noun *argument*, the fourth sense of verb *decide*, and the first sense of adjective*free* as provided by the English and Romanian WordNets; for Romanian we also provide the manual translation into English.

data as subjective, since it represents an essay where "you take a position on a debatable topic and attempt to change readers' minds about it. The more persuasive your argumentative essay, the more likely readers will be to concede your points and grant your conclusion."[5] Instead, the Romanian gloss and synset for this word denote a "direct summary," which by definition disallows the expression of any subjective perspective. Therefore, in Romanian this sense is objective.

A similar scenario is posed by the fourth sense of the verb *decide* (also listed in Table 4.1). While the English sense is labeled as objective, as its meaning denotes causality, the Romanian sense directly implies a subjective decision, and therefore acquires a subjective label.

[5] Writing Literary Arguments - `http://academic.cengage.com/resource_uploads/downloads/1413022812_59427.pdf`

4.1.2. WordNet Granularity

In several cases, the same sense in WordNet may have both subjective and objective meanings. To exemplify, let us consider the first sense of the adjective *free*, as shown in Table 4.1. While the English sense can have both subjective and objective uses, the Romanian sense is subjective, as it further enforces the constraint that the context of the word should refer to people.

From these examples, we notice that a perfect sense to sense mapping among languages is impossible, as a particular sense may denote additional meanings and uses in one language compared to another. While differences or misalignment of senses also occur between monolingual dictionaries within the same language, it is important to remember that the Romanian WordNet was built to be sense aligned with its English counterpart; thus these differences are caused by cross-lingual transfer. However, in this annotation study about 90% of the senses maintained their subjective meaning across languages, implying that this information can be leveraged in an automatic fashion to provide additional clues for the subjectivity labeling of unseen senses.

4.2. Multilingual Subjectivity Sense Learning

In an earlier work exploring the ability of multilingual models to better capture subjectivity at the sentence level (see Chapter 6), which was conducted on six languages, namely English, Arabic, German, Romanian, Spanish and French, it was shown that simultaneously considering features originating from multiple languages results in error rate reductions ranging from 5% for English to 15% for Arabic, when compared to the monolingual model baselines. The experiments also pointed out that the maximum improvement is achieved when the multilingual model is built over the expanded feature space comprising the vocabulary of all six languages. This observation became the catalyst for the work presented in this chapter, as it seeks to explore whether the task of sense level subjectivity classification can also benefit from being modeled with a multilingual perspective in mind, and compare it to a monolingual baseline.

Thus, this chapter explores potential ways to use a multilingual learning mechanism to automatically predict the subjectivity of a word sense, experimenting with two methods. The first one is based on cross-lingual training using monolingual feature spaces. This method uses the output of individually trained monolingual classifiers paired with a set of constraints to reach an overall decision. The second method introduces a learner that is trained on a multilingual feature space, and whose decision is automatically inferred. Ultimately, this scenario is intended to identify whether a classifier is able to make a better decision by having access to the entire feature set.

I start by considering the intersection of the Romanian and English WordNets, so that equivalent senses (including their definitions and synsets) can be extracted in both languages; this resulted in a total of 19,124 unique synsets that are accompanied by a gloss (the synset intersection is substantially higher, but we considered only those synsets that also had a definition in Romanian). Vectorial representations for two monolingual models (one in English and one in Romanian) (see Example 4), and one multilingual model (comprising both Romanian and English features) (see Example 5) are then generated. These are composed of unigrams extracted from a synset and its gloss, appended with a binary weight. The synset is stripped of any sense identifying features[6] in order not to favor the classifier. To exemplify, we provide below the sparse vector representation of the fourth sense of the noun *argument* (see Table 4.1 for its original gloss and synset in English and Romanian):

English vector: <a$_{en}$ 1, summary 1, of 1, the 1, subject 1, or 1, plot 1, literary 1, work 1, play 1, movie 1, editor 1, added 1, argument 1, to 1, poem 1>
Romanian vector: <redare 1, prezentare 1, pe 1, scurt 1, scrisa 1, orala 1, a$_{ro}$ 1, ideilor 1, unei 1, lucrari 1, ale 1, expuneri 1, etc 1, rezumat 1> (4)

Multilingual vector: <a$_{en}$ 1, summary 1, of 1, the 1, subject 1, or 1, plot 1, literary 1, work 1, play 1, movie 1, editor 1, added 1, argument 1, to 1, poem 1, redare 1, prezentare 1, pe 1, scurt 1, scrisa 1, orala 1, a$_{ro}$ 1, ideilor 1, unei 1, lucrari 1, ale 1, expuneri 1, etc 1, rezumat 1> (5)

Traditionally, the subjectivity content of an entity, be it word, sentence, or document, is regarded as a binary decision (either subjective or objective). However, due to the inherent overloading of words in discourse, this section will mimic the natural language usage by using a continuum representation, where 0 is at one end of the spectrum and represents full objectivity, while 1 is at the other end, and denotes full subjectivity. A zone of 0.4 from the left and right of the spectrum is

[6] We only keep the lemma for the words in the synset when we add them to the vectorial representation of a given sense; we do not include any information on the part-of-speech or sense number.

established, and synsets whose scores fall in these ranges are considered as objective (if below 0.4) or subjective (if above 0.6). This allows for the existence of a buffer zone found between the 0.4 and the 0.6 thresholds, that groups those samples that may be considered too vague to be clearly labeled for subjectivity. Because a typical classification approach does not lend itself to being employed under a gradient subjectivity content paradigm (unless mapping the numeric classification scores to nominal buckets), I opted to use a linear regression algorithm as the machine learner[7] which extrapolates from the data and infers a subjectivity score for every synset.

4.2.1. Cross-lingual Learning

The first method focuses on cross-lingual learning (CL). Based on the co-training algorithm proposed by [74], the manually annotated training data in each of the languages is considered individually (as in Example 4), and two monolingual regression algorithms (see Figure 4.1, (1)[8]) are learned. For every sample in the unlabeled data (2), the machine learners predict a score individually (3), and at every iteration two sets with the top n most confident objective and subjective examples, respectively, are maintained. These sets are ordered based on the average calculated between the predictions coming from the English and Romanian learners, which must also fall within the same range (i.e. both below 0.4 or both above 0.6), thus signaling that both learners agree. As long as the sets are not empty (4), at the next iteration the monolingual English vectors and the aligned Romanian vectors are added to their respective training set (+) appended with their adjusted subjectivity score, and removed from their respective test set (-); otherwise the bootstrapping terminates.

Although the method differs from the original co-training mechanism proposed by [12], since it enforces that both predictions fall in the same range before adding the samples to the next train set, this was a necessary modification given the extremely short contexts available, and the low accuracy attained by the English and Romanian classifiers by themselves (67.66% and 70.28%, respectively). Through this additional agreement constraint, only samples that have a high probability of being labeled correctly are added, therefore reducing noise propagation across

[7] Included with the Weka machine learning distribution [23]

[8] The numbers or symbols between parentheses refer to the indices included in the figures.

iterations. At the same time, new information is learned from the features co-occurring with those that participated in the previous classification step.

4.2.2. Multilingual Learning

The second method employs multilingual learning (ML) (see Figure 4.2). Instead of using the monolingual vectors described above, the feature space is enriched by merging two vector space representations that are sense aligned (see Example 5), thus allowing the system to simultaneously use both Romanian and English features in order to decide the subjectivity of a given sense. A multilingual learner is trained (1) and for every sample in the testing set (2), a subjectivity score is predicted (3). Similarly to the cross-lingual learning setup, at every iteration only the most confident n objective and n subjective samples are selected (4) and added to the training set at the end of the iteration (+), while being discarded from the test set to be employed at the next iteration (-).

For both methods, the score of the new samples that are added to the train set during each iteration is mapped to either 0 (objective) or 1 (subjective), the determination being made based on the range in which the original score fell (i.e. if an instance initially received a score of 0.3, since it falls in the objective range its adjusted score will be 0, and the instance will be added to the next iteration training set with this score) . This allows all the training samples to equally participate in the decision process at every iteration, instead of their novel features being penalized due to being absent from the initial training step. For the experiments, 20 iterations were conducted for both methods and 50 subjective and 50 objective samples were added at each iteration. Additional iterations would have been possible, but given the drop in performance of the Romanian learner embedded in the cross-lingual model, the bootstrapping was stopped.

4.2.3. Datasets

From the manually annotated data described in Section 4.1, 20 examples that were labeled by the annotators as both objective and subjective (*both*) in either English or Romanian were removed, since they could confuse the classifiers and prevent them from making strong predictions. However, those synsets that had conflicting labels across languages were maintained in the dataset.

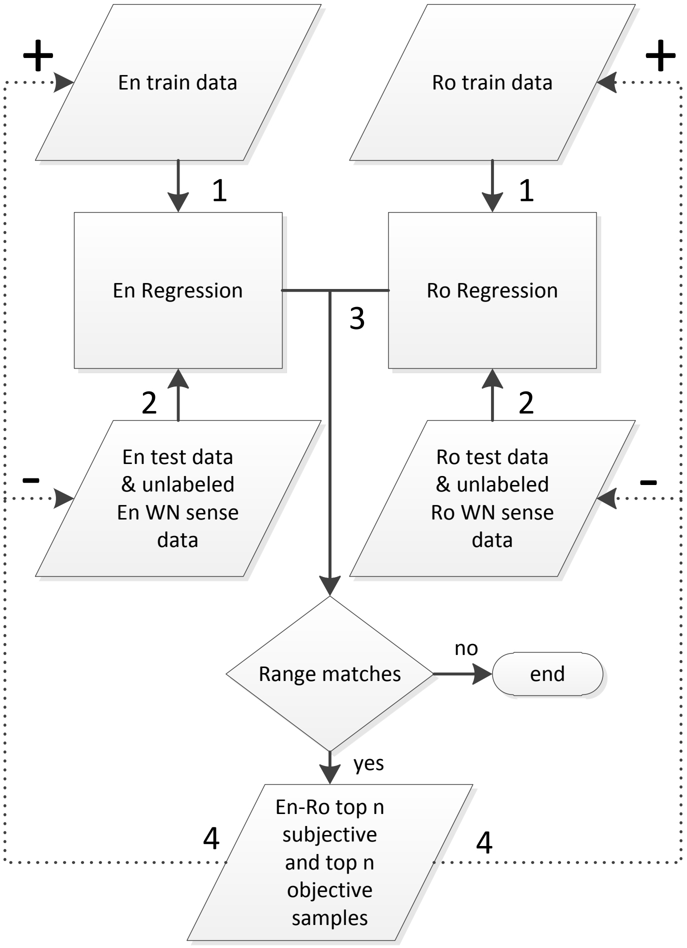

+
En train data
Ro train data
+
1
1
En Regression
3
Ro Regression
2
2
-
En test data & unlabeled En WN sense data
Ro test data & unlabeled Ro WN sense data
-
Range matches
no
end
yes
4
En-Ro top n subjective and top n objective samples
4

FIGURE 4.1. Sense level cross-lingual bootstrapping.

To enable three-fold cross validation, the labeled data is split into three subsets; as these are biased towards the objective class in a ratio of 2:1, about half of the objective samples are randomly discarded, in order to obtain balanced training folds, which allows the experimental setup to not be skewed towards any of the classes. Throughout every iteration, this class balance is maintained, as an equal number of the strongest subjective and objective samples are added to the next train set. Note that the test sets were not balanced. Each fold comprises an initial *train set* of 328 samples and a *test set* of 164 samples, on which the evaluations for the respective fold are carried out. In order to generate a *running test set* for each fold, the remaining unlabeled WordNet senses are appended to each test fold (see Figures 4.1 and 4.2, (2)). These running test sets are used to provide the learners with novel samples (and features) throughout the bootstrapping process. Only 328 training examples were exploited because there is a very limited amount of subjectivity data

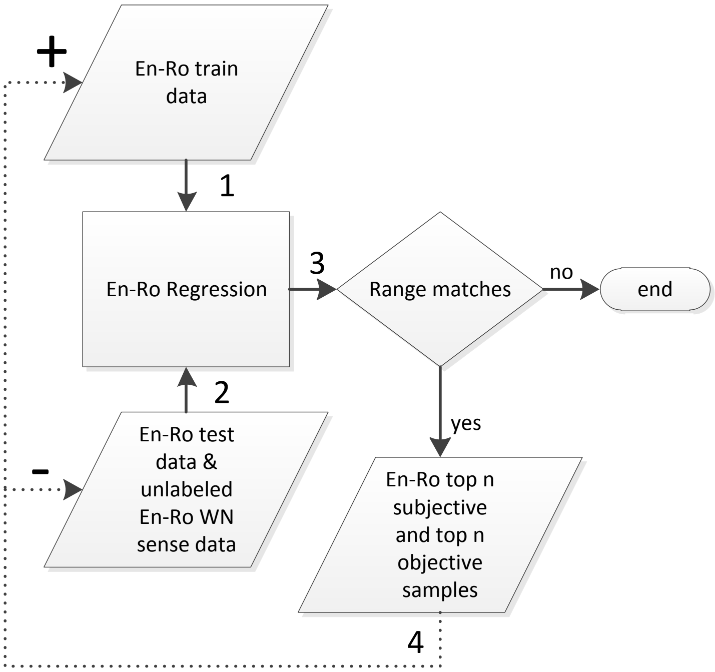

+
En-Ro train data
1
En-Ro Regression
3
Range matches
no
end
2
yes
-
En-Ro test data & unlabeled En-Ro WN sense data
En-Ro top n subjective and top n objective samples
4

FIGURE 4.2. Sense level multilingual bootstrapping.

manually annotated at the sense level in English, which moreover, needs to be mirrored in the Romanian WordNet – which exhibits far less coverage (and thus lower overlap) vis-à-vis the English WordNet. A similar issue will be encountered by most (if not all) of the WordNets developed for languages other than English. From the 580 manually annotated English senses, approximately 500 had an equivalent in the Romanian WordNet. For this reason, the experiments sought to use all the available training examples for both proposed methods as well as the baseline, while also allowing for the existence of a small test set that could be used for evaluation purposes.

4.2.4. Results and Discussions

For the subsequent evaluations, the accuracy and F-measure are calculated based on the score predicted by the linear regression algorithm for every test sample. If the score is higher than 0.5, the sample is considered to belong to the subjective class, otherwise it belongs to the objective class (thus we predict a label for each instance of the test data). The subjectivity continuum described in Section 4.2 is only used internally by the cross-lingual / multilingual bootstrapping methods, since its principal aim is to reduce noise propagation across iterations.

Figure 4.3 presents the results obtained using the cross-lingual learning algorithm over 20 iterations. The accuracies obtained at position 0 represent the baseline for a simple monolingual classifier with no co-training. As noticed from both trendlines, the accuracy for the first 17 iterations is always higher or within 0.56% of the baseline. After the 17th iteration, The Romanian learner drops fast in accuracy loosing 3.72% over the last three iterations, however, the English learner maintains its robustness and in the last iteration is still 2.43% over the baseline. This implies that learners in each of the languages are able to build upon one another and strengthen their prediction, compared to the monolingual scenario; furthermore, they are able to lessen the effect of noise generated at each run, being able to label 1,700 additional test samples (representing more than five times the original training set) with over 69% accuracy in both languages. It is interesting to note that the Romanian learner outperforms the English one throughout all but the last three iterations; a similar trend was noticed when carrying machine learning subjectivity experiments at the sentence level in English and Romanian [5], which was hypothesized to be caused by overt markers of formality and politeness, inflections due to verb mood, and noun and adjective number, gender, and case available in Romanian. The results seem to further support the hypothesis that subjectivity analysis is an easier task in Romanian proposed by [5]. The highest joint accuracy is obtained during the 4th iteration, and it represents a 3.54% improvement over the baseline.

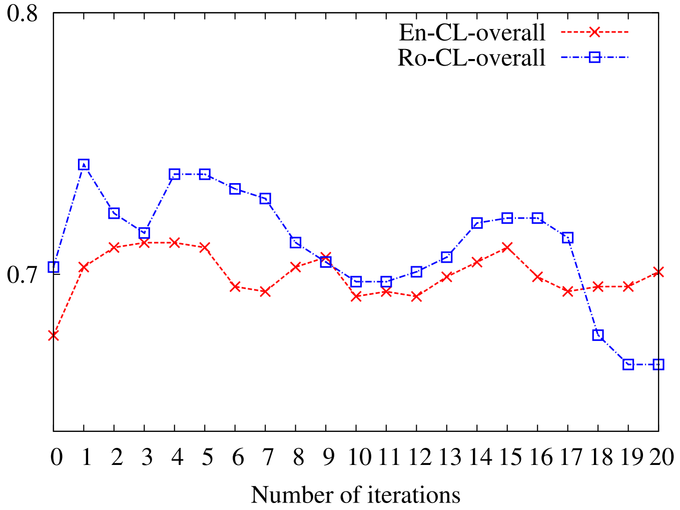


FIGURE 4.3. Macro-accuracy for sense level cross-lingual bootstrapping.

When analyzing the class behavior (see Figure 4.4), the objective samples are more correctly predicted by both learners (an F-measure range from 72% to 78%), when compared to the subjective ones (falling in the 59% to 69% range) irrespective of the underlying language. This is probably the case because glosses and synsets are generally short, and as objectivity is defined through the absence of subjectivity, shorter contexts have a lower probability of containing the manifestation of a private state in comparison to longer ones.

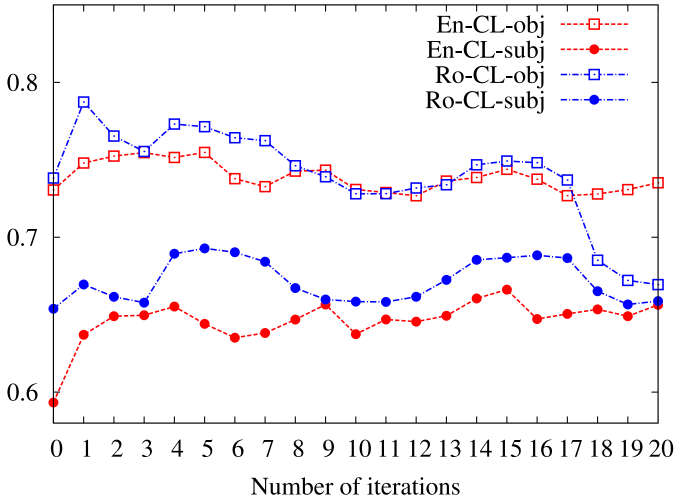


FIGURE 4.4. Per class objective and subjective F-measure for sense level cross-lingual bootstrapping.

In order to understand whether a multilingual vectorial feature space allows for better automatic classification decisions when compared to those taken as a result of heuristics or rules (such as cross-lingual training), a similar experiment is conducted, this time on multilingual vectors. In this scenario, the linear regression algorithm is able to predict a score directly. As seen in Figure 4.5, the multilingual based learner surpasses the cross-lingual based algorithm in all 20 iterations for both languages. Instead of having access to only a fragmented view (as the cross-lingual individual learners use only a monolingual space to make a decision), the multilingual learner has access to the entire feature space, which it uses more proficiently to model subjectivity and thus makes better predictions. Thus, if the baseline for subjectivity classification was 67.66% for English and 70.28% for Romanian, upon having access to the merged feature space, the accuracy for both of them increases to 73.98% (before any iteration takes place), which represents an improvement of 6.32% for English and 3.7% for Romanian. Upon allowing the cumulative effect of this modeling to echo through the iterations, the best results are noticed in iterations 3 and 4, at over 77% accuracy for both languages. Furthermore, the multilingual model is more robust, as it is able to sustain an accuracy of 73% even after 20 iterations, unlike the Romanian cross-lingual learner, which drops to 66.55% in the last iteration.

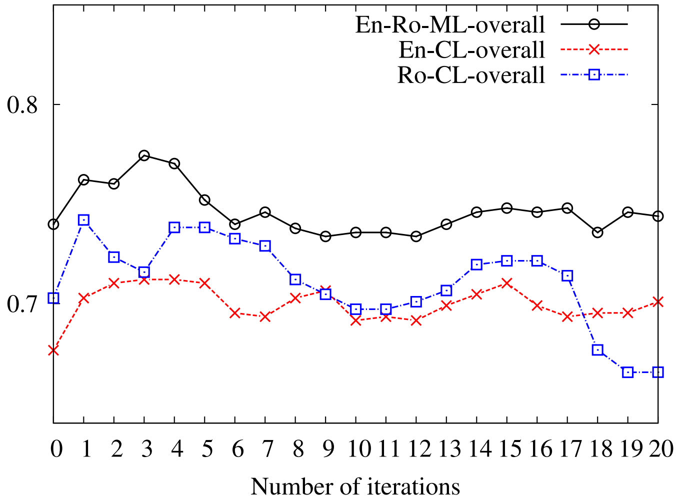


FIGURE 4.5. Macro-accuracy for sense level multilingual bootstrapping compared with the cross-lingual framework.

The class behavior under the multilingual settings is reflected in Figure 4.6. Both the subjective and the objective F-measures are higher than their corresponding F-measures obtained for either English or Romanian. Furthermore, the subjective F-measure increases to over 70% across all the iterations, while the objective one is always higher than 75.6%. In iterations 3 and 4, the objective F-measure is 80%, while the subjective one is 72.7%. Note that while improvement is experienced for both the objective and subjective classes, a major gain is observed for the latter.

I am not aware of any other work that considered the task of word sense subjectivity labeling in a cross-lingual setting, and thus no direct comparison with previous work can be performed. The work closest to ours is the subjectivity word sense disambiguation method proposed in [1], where on a set of 83 English words, an accuracy of 88% was observed; and the method proposed in [62], where an accuracy of 84% was obtained on another dataset of 298 words. These results are however not directly comparable to ours, as they are applied on different datasets, with different levels of difficulty.

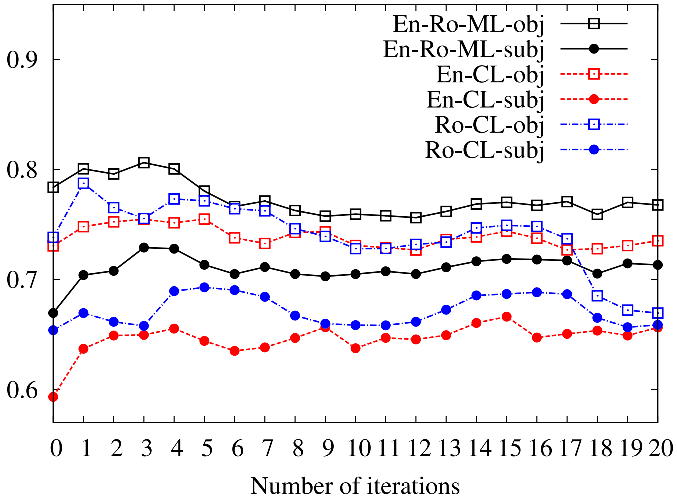


FIGURE 4.6. Per class subjective and objective F-measure for sense level multilingual bootstrapping (versus cross-lingual framework).

I further conduct a qualitative study by applying feature selection based on information gain,[9] and keeping the top 100 features (the study was conducted on the third fold training set generated after the tenth iteration). The subjective entries are listed in Table 4.2 in order of appearance and they show several interesting trends.

First, among the monolingual attributes, the Romanian feature space allows for a more robust selection of subjective words, when compared to the fewer subjective entries in the English set. This is particularly surprising because Romanian is a highly inflected language, and a larger unlemmatized corpus would be needed to extract similar co-occurrence patterns when compared to English. However, this particularity was previously signaled computationally in Figure 4.6, where the subjective F-measure for Romanian is always higher than the subjective F-measure for English.

Second, when looking at the multilingual attributes, we notice that approximately 33% of them are translations of each other (marked in italics in Table 4.2). This shows that the multilingual feature space is able to rely on double co-occurrence metrics learned from equivalent sense definitions, thus allowing a stronger and more accurate prediction to form. This fact is also noticed

[9] Implementation included in the Weka machine learning distribution [23]

Monolingual Features		Multilingual Features
English	**Romanian**	**English & Romanian**
feeling	sentiment (En, n.: feeling)	*feeling*
state	stare (En, n.: state)	*sentiment*
quality	lipsă (En, n.: lack)	mai (En, r.: more)
mental	atitudine (En, n.: attitude)	*lipsă* (En, n.: lack)
feel	suferinţă (En, n.: suffering)	*state*
emotion	boală (En, n.: disease)	not
emotional	simţi (En, v.: feel)	să (subjunctive mood particle)
pain	idee (En, n.: idea)	atitudine (En, n.: attitude)
no	anumit (En, a.: certain)	*stare* (En, n.: state)
hit	sufletească (En, a., fem.: pertaining	good
good	to the soul)	quality
mind	interes (En, n.: interest)	no
great	înţelege (En, v.: understand)	mental
self	părere (En, n.: opinion)	diferite (En, a., fem., pl.: different)
regard	morală (En, a.: moral)	*feel*
important	satisfacţie (En, n.: satisfaction)	*simţi* (En, v.: feel)
judgment	mulţumire (En, n.: contentment)	*lack*
lack	importantă (En, a.: important)	true
true	bună (En, a., fem.: good)	sufletească (En, a., fem.: pertaining
suffering	părăsi (En, v.: abandon)	to the soul)
lacking	provocată (En, a., fem.: provoked)	*pain*
opinion	nelinişte (En, n.: turmoil)	regard
statement	probleme (En, n., pl.: problems)	*suferinţă* (En, n.: pain)
trait	stimă (En, n.: esteem)	self
disposition	afecţiune (En, n.: affection)	încredere (En, n.: trust)
concern	izbi (En, v.: smash)	înţelege (En, v., 3rd person, sg.: un-
extreme	brusc (En, r.: suddenly)	derstand)
felt	dispoziţie (En, n.: mood)	trait
distress	starea (En, n., determined: state)	important
social	să (subjunctive mood particle)	dorinţă (En, n.: desire)
pleasure	calitate (En, n.: quality)	lacking
intense	înţelegere (En, n.: understanding)	
belief	tulburare (En, n.: perturbation)	
danger	simt (En, v., 1st person, sg.: feel)	
feelings	acord (En, n.: agreement)	
argument	durere (En, n.: pain)	
personal	valoare (En, n.: value)	
attitude	emoţie (En, n.: emotion)	
	agitaţie (En, n.: agitation)	
	respect (En, n.: respect)	
	încredere (En, n.: trust)	
	necaz (En, n.: misfortune)	
	spirit (En, n.: mind)	
	însuşire (En, n.: trait)	

TABLE 4.2. Sample of subjective features appearing in the top 100 discriminant attributes selected with information gain on the 3rd fold training data at iteration 10. The words in italics in the rightmost column represent overlapping translations in English and Romanian. Abbreviations: n. - noun, v. - verb, a. - adjective, r. - adverb, fem. - feminine gender, sg. - singular / pl. -plural (in terms of verb person or noun or adjective number).

in Figure 4.6, where the multilingual regression model surpasses the monolingual one by 6.4 subjective F-measure percentage points on average for English, and 3.8 for Romanian, respectively (the average is computed across all folds and iterations).

Third, more than half of the top 100 features obtained as a result of filtering the monolingual and multilingual models using information gain are not subjective from a human annotator's point of view. This shows that the regression algorithm relies on objective markers, thus explaining the improved performance in correctly identifying the objective class, as noticed in Figure 4.6. It is interesting to note that using a multilingual space mainly helps the subjective class, as the objective class average F-measure improves by an average of 3.3% with respect to both monolingual models (the average is computed across all folds and iterations).

Fourth, both the monolingual Romanian space and the multilingual English - Romanian space contain the subjunctive mood particle *să* that is unique to Romanian. Subjunctive is a grammatical verbal mood (part of the irrealis moods), used to mark ideas that are subjective or uncertain, such as emotions, doubts, opinions, judgments, etc., and it provides a unique marker for subjectivity. While in English subjunctive is not distinctively identified, in Romanian a verb in subjunctive is always accompanied by the *să* particle. A discussion regarding this aspect is included in Chapter 7. The particle appears in the ranked Romanian attribute selection list in position 75, yet upon learning from the multilingual space it becomes a highly distinguishing feature, earning position 29. This represents one unique example of a way in which a language provides valuable input to accurately classifying subjectivity in another language.

This case study provides evidence that a multilingual feature space representation of subjectivity at the sense level allows for a more robust modeling than when considering each language individually. Subjectivity clues seem to be able to permeate from each language and simultaneously participate in joint decisions, thus making stronger and more accurate predictions. As private states tend to remain the same across languages (see the manual annotation study in Section 4.1), these results strengthen the hypothesis that subjectivity is a language independent phenomenon, and as such, it can only gain a stronger contour when considering its emergence from across a number of languages.

While I was not able to conduct a study on the difference in performance when using different language pairs for sense subjectivity annotations, mainly because of not having the required sense resources in other languages, however, previous work on subjectivity at sentence level (see Chapter 6) seems to indicate that a more robust learning occurs when the languages are further apart. In that scenario we learned subjectivity from up to 6 languages (English, German, Arabic, Spanish, French and Romanian) at a time, and it was interesting to note that in all combinations of six languages taken 2 through 4, English did not participate in the top performing combination (as it did in the monolingual model). Instead, it got replaced by the space generated by German and Spanish, which offered a better model for subjectivity.

4.3. Conclusion

The case study presented in this chapter sought to assess subjectivity transfer across languages following sense aligned resources. In the annotation experiments the subjectivity content of a sense carried across language boundaries in about 90% of the cases, implying that this information is robust enough to be learned automatically. I then proposed and applied a framework that is able to jointly exploit the subjectivity information originating from multiple languages. The machine learning experiments demonstrate that learning sense level subjectivity from a multilingual feature space is able to capture more information and outperform cross-lingual learning based on monolingual vectorial models, while also allowing for even better results to be obtained upon bootstrapping.

CHAPTER 5

WORD AND PHRASE LEVEL MULTILINGUAL SUBJECTIVITY

Numerous approaches to automatic sentiment and subjectivity detection rely on subjectivity lexica, namely words and phrases that are endowed with a subjective component (i.e. have a subjective usage). These can be constructed either manually, such as the General Inquirer lexicon [56] or through an automatic process [85, 52, 34]. These resources are particularly useful, as they can be employed in rule-based approaches that calculate scores reflecting the presence of subjectivity clues (e.g. [87]), or in machine learning approaches that use features and / or weights based on pre-existing lexicon entries [52, 83].

In light of the success achieved by techniques for automatic lexicon construction, this chapter seeks to answer the following questions:

Question 1: Assuming a subjectivity annotated English lexicon, can we use machine translation to generate a high quality lexical resource in a target language?

Question 2: If we translate a small number of seeds from this subjectivity lexicon, can we grow the set while maintaining the quality of the lexicon in the target language?

Two potential paths to leverage subjectivity annotated lexicons originating in a source language are explored. One is based on attempting to automatically translate a source language lexicon into the target language by selecting the first sense from a bridging multilingual dictionary (see Section 5.1) . The resources required by this experiment are an annotated subjectivity lexicon and a multilingual dictionary. The second is based on selecting a small set of subjective seeds from the source language lexicon and manually translating them into the target language (see Section 5.2). We can then expand the lexicon by solely using material in the target language through a bootstrapping mechanism. This scenario requires the availability of a small set of subjective seeds, an electronic dictionary, and a raw corpus. Both methods are evaluated by building a rule-based classifier that relies on the generated lexicons to perform subjectivity annotations in the target language.

Parts of this research were previously published in [43, 5, 9].

5.1. Translating a Subjectivity Lexicon

Question 1: Assuming a subjectivity annotated English lexicon, can we use machine translation to generate a high quality lexical resource in the target language?

The OpinionFinder subjectivity lexicon (introduced in Section 3.1.3) acts as a lever in the scenario described above. Since this subjectivity lexicon is compiled based on occurrences in actual text, our first task is to extract words in the form in which they may appear in a dictionary. Due to the fact that some of the entries present discrepancies, we implement a voting mechanism that takes into consideration the additional annotations such as part-of-speech, polarity, etc., accompanying every entry, to decide on the correct part-of-speech information and lemma. In case of a tie, we query WordNet to decide which form to consider. To exemplify, let us consider the word “atrocities.” For this entry, the *origpats* field suggests an erroneous adjective classification, further supported by the *pos1* field. The voting mechanism is however able to accurately decide on the correct part-of-speech, by taking into consideration the *othertypes*, *stem1*, *origtypes*, *origpats*, *stemmed1*, and *type* fields. Therefore, this subjective entry is corrected, and its lemma becomes “atrocity,” while its part-of-speech is updated to *noun*.

```
′atrocities′ = { ′len′ => ′1′,
   ′word1′ => ′atrocities′,
   ′othertypes′ => ′metaboot-2000nouns-strongsubj′,
   ′pos1′ => ′adj′,
   ′highprec′ => ′yes′,
   ′polannsrc′ => ′ph′,
   ′stem1′ => ′atrocity:noun′,
   ′origtypes′ => ′metaboot-2000nouns-strongsubj′,
   ′RFE′ => ′tff′,
   ′origpats′ => ′%atrocities||adj||n%′,
   ′mpqapolarity′ => ′strongneg′,
   ′stemmed1′ => ′n′,
   ’MISS’ => ’t’,
   ’intensity’ => ’high’,
   ’type’ => ’metaboot-2000nouns-strongsubj’ };
```

In the case of “loathing” (see the example below), the *origpats* (part-of-speech in the original context) field suggests that the correct part-of-speech is verb, yet when looking at the *stem1* field, the possibilities are either verb or noun. Since the lemma of the word would have a different form based on the selected part-of-speech, the voting schema corrects this entry as well; a vote for verb is cast by *origpats*, *stem1*, *type*, and *origtypes* fields, while a vote for noun is only cast by *stem1*. We therefore conclude that the correct part-of-speech for loathing is *verb*, and its lemma is *loathe*.

```
’loathing’ = { ’len’ => ’1’,
     ’origpats’ => ’%loathe||verb||y%, %loathe||verb||y%’,
     ’word1’ => ’loathing’,
     ’mpqapolarity’ => ’strongneg’,
     ’stemmed1’ => ’n’,
     ’pos1’ => ’anypos’,
     ’polannsrc’ => ’ph’,
     ’type’ => ’fn_emotion_experiencer-subj_v’,
     ’polarity’ => ’negative’,
     ’intensity’ => ’high’,
     ’stem1’ => ’loathe:verb#loathing:noun’,
     ’origtypes’ => ’bl_psych_verb,fn_emotion_experiencer-subj_v’,
     ’RFE’ => ’fff’ };
```

Once the discrepancies are adjudicated, the lexicon is filtered for entries composed of a single word, and which are further labeled as either strong or weak (this information is provided by the *mpqapolarity* field). This results in a finalized version of the subjectivity lexicon in the

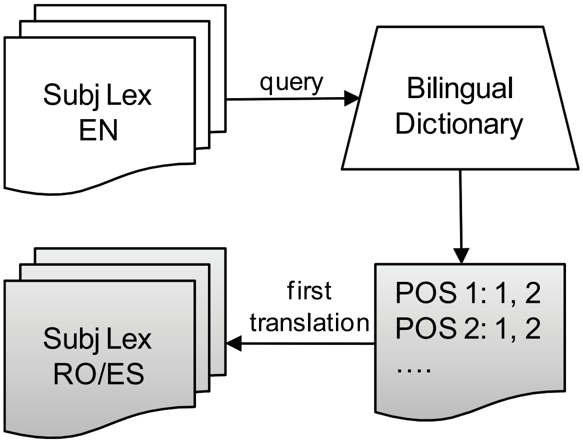


FIGURE 5.1. Automatic translation process of a subjectivity lexicon from source language into target language.

source language, containing 5,339 entries. Next, the Ectaco online dictionary[1] is queried for each entry in order to perform the translation in the target language. This service provides translations into 25 languages, and each dictionary features more than 400,000 entries. We chose this resource in order to be able to conduct a quality assessment of our approach, by carrying out translation experiments from the source language into two languages that are both supported by the Ectaco online dictionary, namely Romanian and Spanish.

The automatic source language lexicon translation is exemplified in Figure 5.1.

While the task of translating a lexicon might seem trivial at a first sight, there were several challenges encountered in the translation process. First, although the English subjectivity lexicon contains inflected words, we must use the lemmatized form in order to be able to translate the entries using the bilingual dictionary. However, words may lose their subjective meaning once lemmatized. For instance, the inflected form of "memories" becomes "memory." Once translated into Romanian (as "memorie"), its main meaning is objective, referring to the power of retaining information as in: "Iron supplements may improve a woman's memory."[2]. Therefore it is very difficult if not impossible to recreate the inflection, as Romanian does not have a synonym for "memories" from the same lexical family.

[1] http://www.ectaco.co.uk/free-online-dictionaries/

[2] The correct translation in this case for the plural "memories" would have been "amintiri"

Second, neither the lexicon nor the bilingual dictionary provide information on the sense of the individual entries, and therefore the translation has to rely on the most probable sense in the target language. Fortunately, the bilingual dictionary lists the translations in reverse order of their usage frequencies. Nonetheless, the ambiguity of the words and the translations still seems to represent an important source of error. Moreover, the lexicon sometimes includes identical entries expressed through different parts of speech, e.g., "grudge" has two separate entries, for its noun and verb roles, respectively. On the other hand, the bilingual dictionary may not make this distinction, and therefore we may have again to rely on the *most frequent* heuristic captured by the translation order in the bilingual dictionary.

Third, Romanian and Spanish do not offer direct translations for the multitude of adverbs suffixed by *-ly* in English (e.g., the adverb "freely" obtained from the noun "free" can be translated into Romanian or Spanish only as a phrase (Ro: "în libertate", Es: "con libertad"; En: "in/with freedom")). Others, such as "staunchly" or "solicitously" do not return any translations.

The Ectaco dictionary provided similar coverage for both target languages, as 1,580 entries were translated into Romanian (29.6%) and 2,009 (37.6%) into Spanish, respectively. Table 5.1 shows examples of entries in the Romanian and Spanish lexicons, together with their corresponding original English form. The table also shows the reliability of the expression (*weak* or *strong*) and the part-of-speech, both attributes being provided by the English subjectivity lexicon.

English	attributes	Romanian	Spanish
admonish	strong, verb	preveni	amonestar
beautify	strong, verb	împodobi	embellecer
credence	weak, noun	crezare	creencia
diligent	strong, adj	sârguitor	aprovechado
excuse	weak, verb	scuză	disculpa

TABLE 5.1. Examples of entries in the Romanian and Spanish subjectivity lexicon.

5.1.1. Manual Evaluation

It is important to assess the quality of the translated lexicons, and compare them to the quality of the original English lexicon. The English subjectivity lexicon was evaluated in [80] against a corpus of English-language news articles manually annotated for subjectivity. According to that evaluation, 85% of the instances of the clues marked as *strong* and 71.5% of the clues marked as *weak* appear in subjective sentences in the MPQA corpus [80], where 55% of the sentences in this corpus are subjective.

Since there are no comparable Romanian or Spanish corpora, an alternate way to judge the subjectivity of the translated lexicon entries is needed. Two native speakers, one of Romanian and one of Spanish, annotated the subjectivity of 150 randomly selected entries in the generated lexica. They were presented with the original English subjective entry and the automatically translated one. Due to word ambiguity, and the inability of a human to immediately recall all the possible senses and uses of a given word, the judges were helped by also being provided with the first approximately 100 snippets containing the translated word, based on a query to the Google search engine (restricted to either Spanish or Romanian). Since many of the sites in languages with fewer electronic resources provide news, and their content changes more frequently, thus influencing the Google search results ranking, a large number of snippets originates from the news domain. The subjectivity of a word was consequently judged from the contexts in which it most frequently appeared, thus accounting for its most frequent meaning on the Web. The tag set used for the annotations consists of *Subj(ective)*, *Obj(ective)*, and *Both*[3]. A *Wrong* label is also used to indicate a wrong translation. Additionally, for the *Subj(ective)* and *Both* labels, the judges added strength granularity, resulting in *weak* and *strong* annotations. An entry is considered *strong* when its appearance in a given context would render the entire text subjective. In contrast, a *weak* entry contains a certain level of subjectivity, yet a sentence in which a weak entry appears may not be labeled as subjective based on this clue alone. Table 5.3 summarizes the two annotators' judgments on this data.

Since the English subjectivity lexicon does not provide explicit annotations for cases where an entry may have both subjective and objective uses (*Both*), and since this label was part of the tag set, the annotation study was repeated in English on the same 150 randomly selected entries. Two English native speakers annotated the entries according to the same guidelines used for the Romanian and Spanish annotation studies. At the end of the annotation session the judges discussed

[3] *Both* is used when the word does not have a clear subjective or objective predominant use, but can rather appear in both types of contexts equally.

their disagreements and decided on a gold-standard for the entire word set. A comparison between the annotations extracted from the OpinionFinder lexicon and the one we conducted is presented in Table 5.2. It is interesting to note that all the entries that were labeled as *Both* represent weak indicators of subjectivity. We were only able to calculate the subjectivity *strength* agreement between the initial labels and those resulted after the study. The agreement is 0.78; due to the skewed distribution of the entries towards *strong* subjectivity, Kappa is κ=0.50, which indicates moderate agreement [37].

If the original annotations for the English subjectivity lexicon classified 65.33% (98) of the entries as strong and 34.67% (52) as weak, as a result of the additional annotation study, 8.66% (13) of the entries had both objective and subjective uses, while 91.34% (137) were labeled as subjective; 68.67% (103) of the entries were labeled as strong and 22.67% (34) as weak.

OpinionFinder Lexicon		Manual Annotation Study			
Strong	Weak	Strong		Weak	
65.33% (98)	34.67 (52)%	68.67% (103)		31.33% (47)	
		Subj 68.67 (103)%	Both 0% (0)	Subj 22.67% (34)	Both 8.66% (13)

TABLE 5.2. English annotation study of 150 lexicon entries.

Thus, the study presented in Table 5.3 suggests that the Romanian and Spanish subjectivity clues derived through translation are less reliable than the original set of English clues. Only 70% of the translated entries into Romanian and 72% of those translated into Spanish are considered unequivocally subjective by the judges. Also, about 19% of the entries automatically ported to both Romanian and Spanish have ambiguous subjective meanings; out of these, 30% mirror the *Both* tag conferred in the English manual annotation study. It is also interesting to note that the behavior of the two languages is very similar, as they differ by at most three annotations for each tag category.

In several cases, the subjectivity is lost in translation, mainly due to word ambiguity in either the source or target language, or both. For instance, the word "fragile" correctly translates into Romanian as "fragil", yet this word is frequently used to refer to breakable objects, and it loses its subjective meaning of "delicate". Other words completely lose subjectivity once translated

Lang	Subj	Both	Obj	Wrong
Ro	70% (105)	19.33% (29)	6.66% (10)	4% (6)
Es	72% (108)	18% (27)	5.33% (8)	4.67% (7)
En	91.33% (137)	8.66% (13)		
OF_{lex}	100% (150)			

TABLE 5.3. Evaluation of 150 entries in the Romanian (Ro) and Spanish (Es) lexicons, and comparison with the English manual annotation study (En) and the OpinionFinder English lexicon (OF_{lex})

(such as "one-sided," , which becomes "cu o singură latură" in Romanian, meaning "with only one side" (as of objects)). In the case of verb "appreciate" translated into Romanian as "aprecia," which is a polysemous verb denoting a frequent objective meaning of "gaining value" (as of currencies); in Spanish, the word was translated as "estimar" ("estimate"), which involves a far more clear subjective judgment.

In other cases, the translation adds frequent objective meanings through part-of-speech transfer. One example is the adverb "icy," which the dictionary translates into the noun "gheaţă" ("ice") in Romanian; due to the transfer in part-of-speech, the translation candidate has only an objective meaning.

In a similar way, the word *strut* (see definition below) appears in the subjectivity lexicon as a verb. Once the translation is performed, its correspondent in Spanish becomes the noun *puntal*, best defined as the noun *strut* in its first dictionary sense, with a clear objective meaning and use.

strut[4]

intransitive verb

1: to become turgid : swell

2 a: to walk with a proud gait

b: to walk with a pompous and affected air

noun

1: a structural piece designed to resist pressure in the direction of its length

2: a pompous step or walk

3: arrogant behavior : swagger

[4] Definition provided by Merriam Webster online dictionary.

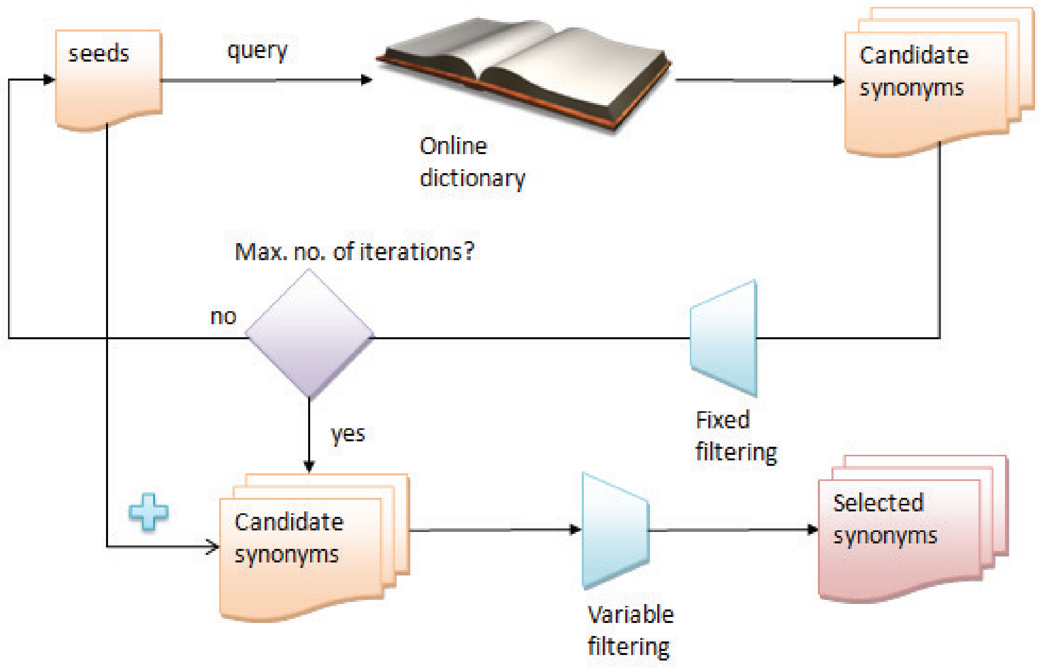


FIGURE 5.2. Bootstrapping process for generating a subjectivity lexicon in a target language.

transitive verb

: to parade (as clothes) with a show of pride

Furthermore, for the verb *boil* the English-Spanish bilingual dictionary proposes *furúnculo*, a direct translation of the noun sense of this word.

Noticing that part-of-speech information bears an important role on the accurate disambiguation of both source and target language words, we also experimented with restraining the translation by enforcing the part-of-speech of the source lexicon entry. Section 5.4 discusses the results obtained following this scenario.

5.2. Growing a Subjectivity Lexicon

Question 2: If we translate a small number of seeds from this subjectivity lexicon, can we grow the set while maintaining the quality of the lexicon in the target language?

A different path to generate a subjectivity lexicon in a target language is to acquire a large subjectivity lexicon by bootstrapping from a few manually selected seeds. At each iteration, the seed set is expanded with related words found in an on line dictionary, which are filtered by using a measure of word similarity. The bootstrapping process is illustrated in Figure 5.2.

5.2.1. Seed Set

I select a preliminary set of 60 seeds, evenhandedly sampled from verbs, nouns, adjectives and adverbs. This number is motivated by the fact that it includes sufficient words in each part-of-speech grouping (approximately 15), thus enabling evaluations under a variety of settings when growing the subjectivity lexicon (*seed vs. candidate*, *POS group vs. candidate*, *all vs. candidate*). While seeds can easily be obtained directly in the target language, without the need of manual translation from a source language, in order to maintain similar experimental settings across languages, I opted to manually translate the seeds into Romanian and Spanish starting from hand-picked strong subjective entries appearing in the English subjectivity lexicon. Table 5.4 shows a sample of the entries in the initial seed set translated into the target languages, accompanied by the initial seed word in English. A similar seed set can be easily obtained for any other language, either by finding a short listing of subjective words in the language of interest or by manually translating a small set of subjective entries from English.

POS	Sample seeds
Noun	blestem / maldiciòn (curse), despot / tirano (tyrant), furie / furia (fury), idiot / idiota (idiot), fericire / felicidad (happiness)
Verb	iubi / amar (love), aprecia / apreciar (appreciate), spera / esperar (hope), dori / desear (wish), urî / odiar (hate)
Adj	frumos / bello (beautiful), dulce / dulce (sweet), urât / feo (ugly), fericit / feliz (happy), fascinant / facinante (fascinating)
Adv	posibil / posiblemente (possibly), probabil / probablemente (probably), desigur / seguramente (of course), enervant / irritante (unnerving)

TABLE 5.4. Sample entries from the initial seed set in Romanian (Ro) /Spanish (Es) accompanied by their English translations.

5.2.2. Dictionary

Starting with the seed set, new related words are added based on the entries found in a dictionary. For each seed word, all the open-class words appearing in its definition are collected, appended with synonyms and antonyms, if available. Since all the possible meanings for each candidate word are expanded and processed, word ambiguity is not an impediment for this method.

In the experiments, for the Romanian dictionary I use *Dex online*,[5] while for the Spanish dictionary I query the *Diccionario de la lengua española*[6] maintained by the Real Academia Española institution. Similar dictionaries are available for many other languages; when online dictionaries are not available, they can be obtained at relatively low cost through OCR recognition performed on a hard copy dictionary.

5.2.3. Bootstrapping Iterations

For each seed word, a query is formulated against the explicative dictionary available in the target language. From the definitions obtained, a list of related words is extracted and added to the list of candidates if they were not already encountered, if they are longer than three characters, and if they do not appear in a list of stopwords.

Three different variants are used to filter the candidate words. The first focuses on capturing the similarity between the original seed word that extracted the candidate and the candidate (*seed vs. candidate*). The second variation groups together all seeds with the same part-of-speech, and proposes calculating the similarity between the candidate and the group with the part-of-speech that extracted it (*POS group vs. candidate*). The third variation filters candidates based on their similarity with the entire original seed set (*all vs. candidate*).[7] The bootstrapping process continues to the next iteration until a maximum number of iterations is reached.

Note that the part-of-speech information is not maintained during the bootstrapping process, as candidate words occurring in the definitions belong to different parts-of-speech and I do not use a POS tagger (since the method should be easily portable to other languages). Although the initial seed set is balanced with respect to syntactic categories, as candidate words are extracted, the balance may be skewed toward one of the categories by the end of the bootstrapping process.

[5] http://www.dexonline.ro

[6] http://buscon.rae.es/draeI/

[7] The intuition behind using P0S groupings is that we will be able to capture functionally similar words which, since they are compared to a subjective set of seeds, will also be subjective. In the case of the 'all' grouping, a stronger measure of subjectivity is expected to emerge, as subjective dimensions end up being added together.

5.2.4. Filtering

In order to remove noise from the lexicon, a filtering step is implemented which is performed by calculating a measure of similarity between the original seeds (in the three variations mentioned earlier) and each of the possible candidates. I experimented with two corpus-based measures of similarity, namely the Pointwise Mutual Information [70] and Latent Semantic Analysis (LSA) [17, 36]. I ultimately decided to use only LSA, as both methods provided similar results, but the LSA-based method was significantly faster and required less training data. The experiments use the Infomap NLP[8] implementation of LSA. After each iteration, only candidates with a LSA score higher than 0.4 (determined empirically) are considered to be expanded in the next iteration.

Upon bootstrapping termination, the subjectivity lexicons constructed incrementally after each iteration consist of a ranked list of candidates in decreasing order of similarity to the three seed set variations. A variable filtering threshold can be used to enforce the selection of only the most closely related candidates, resulting in more restrictive and pure subjectivity lexicons. The following thresholds were used in the experiments: 0.40 (i.e., the lexicon resulting after the bootstrapping process without additional filtering), 0.45, 0.50, and 0.55.

The LSA module was trained on a 57 million word corpora we constructed for Romanian and Spanish. Smaller corpora are also feasible [5], yet in order to obtain a more accurate LSA similarity measure, larger data sets are desirable. Corpora can be obtained for many low-resource languages by using semi-automatic methods for corpus construction [22]. The Romanian corpus was created from a set of editorials collected from Romanian newspapers, a set of publicly available Romanian literature accessible from WikiSource, the Romanian Wikipedia, and the manual translation of a subset of the SemCor data set [44] into Romanian. The Spanish corpus is similar in composition to the Romanian data set, yet it uses one seventh of the Spanish Wikipedia, no editorials, and no SemCor data. The size of this corpus was limited in order to create settings that are as similar as possible to the experiments conducted in Romanian. Also, adding literature works or editorials to the mix is motivated by the need to increase the occurrence of potentially subjective entries in the corpora. Since Wikipedia is considered to be an encyclopedic resource, its usage of

[8] http:infomap-nlp.sourceforge.net

subjective entries may be limited, therefore impairing LSA's ability in calculating similarity among subjective words.

5.3. Gold-Standard

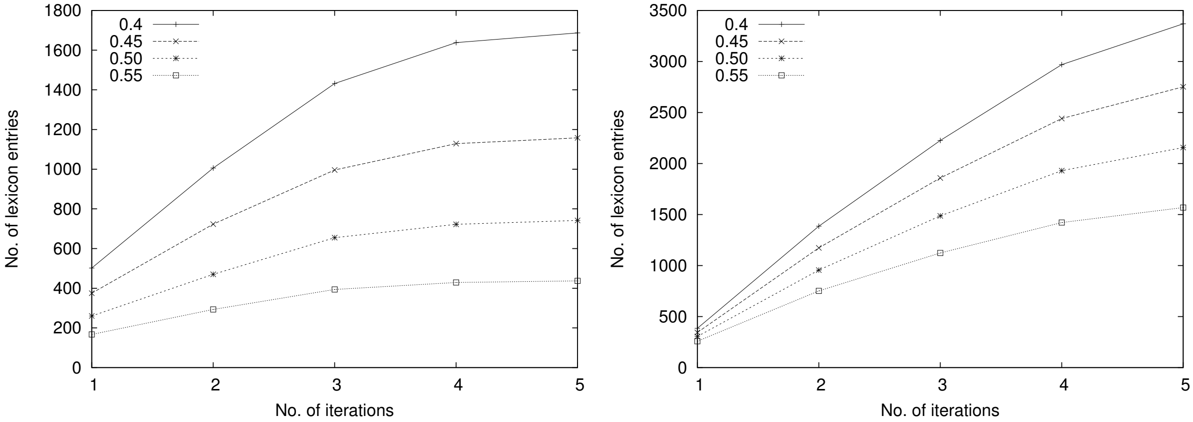
The results are evaluated against a gold-standard consisting of 504 sentences extracted from the English SemCor corpus [44]. These sentences were manually translated into Romanian and Spanish, resulting in two parallel test sets for the two languages. Two Romanian native speakers annotated the Romanian sentences individually, and the differences were adjudicated through discussions. The agreement of the two annotators is 0.83% ($\kappa = 0.67$); when the uncertain annotations are removed, the agreement rises to 0.89 ($\kappa = 0.77$). The two annotators reached consensus on all sentences for which they disagreed, resulting in a gold-standard dataset with 272 (54%) subjective sentences and 232 (46%) objective sentences. The same subjectivity annotations developed for Romanian are also ported to the Spanish test set. The test set is further processed by removing diacritics, and any non literal characters. The corpus based methods described in Chapter 6 use this version of the test sets. For the lexicon based methods proposed earlier in this section, the test data is further lemmatized in order to allow for a match with the automatically extracted candidates from the dictionaries. Lemmatization for Romanian is performed automatically using the module provided by the LanguageWeaver translation engine.[9] For Spanish, the TreeTagger program was used.[10]

5.4. Evaluation and Discussion

The experiments suggest that five bootstrapping iterations are sufficient to extract a subjectivity lexicon, as the number of features saturates during the last iteration. Figures 5.3, 5.4, and 5.5 exemplify the lexicon acquisition in both Romanian and Spanish through all five iterations, using different filtering variations as proposed in Section 5.2.3. As suggested by the graphs, as a stricter similarity between the seeds and the candidates is enforced, a lower number of entries are extracted. Also, the number of entries extracted in Romanian are consistently fewer when compared to Spanish by a factor of at least 1 to 2, resulting in a lower lexicon coverage in this language.

[9] http://www.languageweaver.com

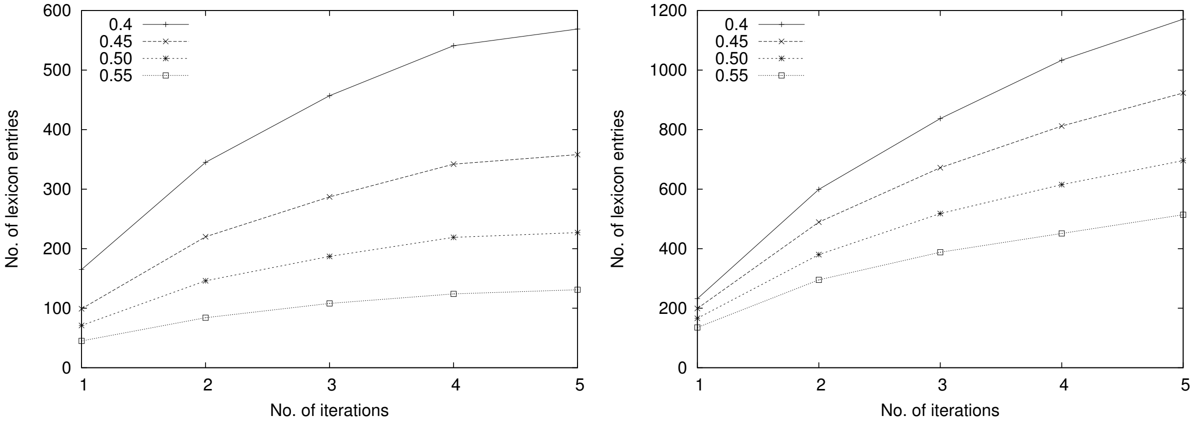
[10] developed by Helmut Schmid as part of the TC project at the Institute for Computational Linguistics, University of Stuttgart



(a) Romanian

(b) Spanish

FIGURE 5.3. Lexicon acquisition over five iterations using the *seed v. candidate* variation.

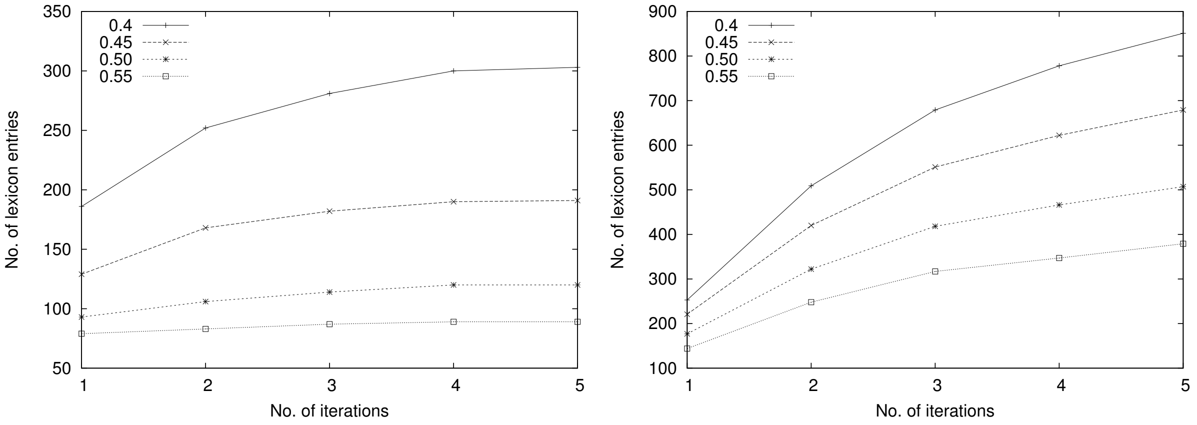


(a) Romanian

(b) Spanish

FIGURE 5.4. Lexicon acquisition over five iterations using the *POS group v. candidate* variation.

The variations proposed for the filtering step during the bootstrapping process resulted in three lexicons for Romanian and Spanish respectively, which are evaluated by using them with a rule-based sentence-level subjectivity classifier. Briefly, the rule-based algorithm labels as subjective a sentence that contains two or more entries that appear in the subjectivity lexicon, and as objective a sentence that has one or fewer entries, respectively. The algorithm is derived based on the rules described in [80], which were modified to account for the fact that no strong/weak confidence labels are available.



(a) Romanian

(b) Spanish

FIGURE 5.5. Lexicon acquisition over five iterations using the *all v. candidate* variation.

The sentence-level subjectivity classification results are shown in Tables 5.5, and 5.6[11]. By using the extracted lexicons alone, a rule-based subjectivity classifier was obtained featuring an overall F-measure of 64.29% (*seed*), 56.94% (*POS group*), and 52.38% (*all*), for Romanian, and 57.54% (*seed*), 64.09% (*POS group*), 69.64% (*all*), for Spanish. The *seed* variation of the bootstrapping process entails a more lax similarity, since it focuses on extracting candidates based on their closeness to an individual seed word, and therefore is able to extract the largest lexicons in the experiments both in Romanian and Spanish (see Figure 5.3). The other two variations gradually enforce a stricter similarity, as the comparison is first made against 15 seeds for the *POS group*, and then against 60 seeds for the *all* set. Each candidate in this case not only has to display similarity with one of the seeds, but with each and every element composing the set. For this reason, the *POS group* variation is able to extract a medium size lexicon (Figure 5.4), while the *all* variation derives the most compact and highly correlated lexicon (Figure 5.5), in both Romanian and Spanish.

As noticed from Tables 5.5 and 5.6 Romanian seems to exhibit an opposite F-measure pattern over the method variations when compared to Spanish. Yet, this is only a superficial assessment. As shown in Figures 5.6 and 5.7, the subjectivity precision and recall curves perform very similarly in both Romanian and Spanish. The only factor that creates a discrepancy is the low

[11] All evaluations are performed with the *Statistics::Contingency* Perl module; the reported *overall* precision, recall and F-measure are micro-averaged.

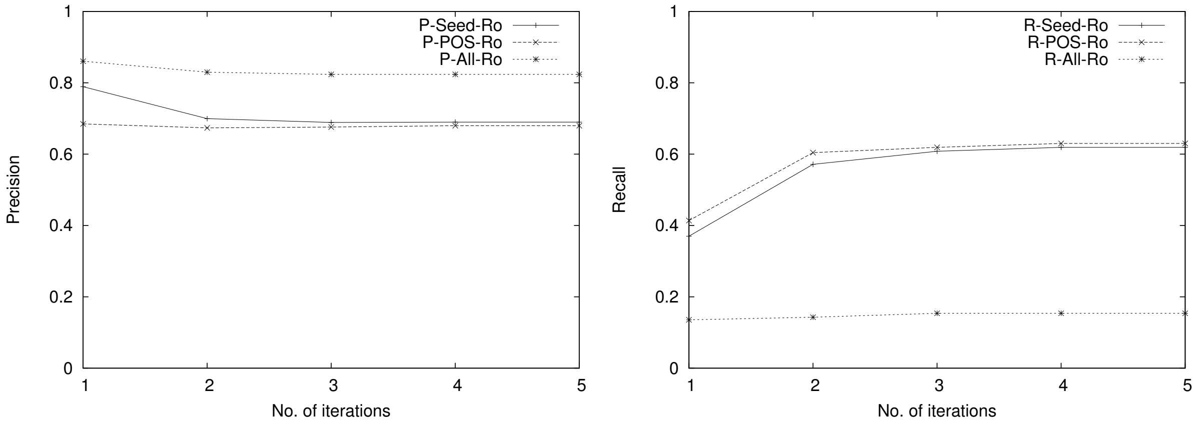
		Seed			POS			All		
Iter.	Eval.	Overall	Subj.	Obj.	Overall	Subj.	Obj.	Overall	Subj.	Obj.
1	P	60.52%	78.91%	54.26%	50.79%	80.49%	48.16%	51.98%	86.05%	48.81%
	R	60.52%	37.00%	88.31%	50.79%	12.09%	96.54%	51.98%	13.55%	97.40%
	F	60.52%	50.37%	**67.22**%	50.79%	21.02%	64.27%	51.98%	23.42%	65.03%
2	P	63.49%	69.96%	58.36%	55.36%	76.09%	50.73%	51.98%	82.98%	48.80%
	R	63.49%	57.14%	71.00%	55.36%	25.64%	90.48%	51.98%	14.29%	96.54%
	F	63.49%	62.90%	64.06%	55.36%	38.36%	65.01%	51.98%	24.38%	64.83%
3	P	63.89%	68.88%	59.32%	56.55%	76.47%	51.49%	52.38%	82.35%	49.01%
	R	63.89%	60.81%	67.53%	56.55%	28.57%	89.61%	52.38%	15.38%	96.10%
	F	63.89%	64.59%	63.16%	56.55%	41.60%	65.40%	**52.38**%	**25.93**%	**64.91**%
4	P	64.29%	68.98%	59.85%	56.94%	76.42%	51.76%	52.38%	82.35%	49.01%
	R	64.29%	61.90%	67.10%	56.94%	29.67%	89.18%	52.38%	15.38%	96.10%
	F	**64.29**%	**65.25**%	63.27%	**56.94**%	**42.74**%	**65.50**%	**52.38**%	**25.93**%	**64.91**%
5	P	64.29%	68.98%	59.85%	56.94%	76.42%	51.76%	52.38%	82.35%	49.01%
	R	64.29%	61.90%	67.10%	56.94%	29.67%	89.18%	52.38%	15.38%	96.10%
	F	**64.29**%	**65.25**%	63.27%	**56.94**%	**42.74**%	**65.50**%	**52.38**%	**25.93**%	**64.91**%

TABLE 5.5. Romanian; Precision (P), Recall (R) and F-measure (F) for the bootstrapping subjectivity lexicon over 5 iterations and an LSA threshold of 0.5

		Seed			POS			All		
Iter.	Eval.	Overall	Subj.	Obj.	Overall	Subj.	Obj.	Overall	Subj.	Obj.
1	P	58.13%	68.02%	53.01%	56.75%	78.35%	51.60%	56.15%	78.26%	51.21%
	R	58.13%	42.86%	76.19%	56.75%	27.84%	90.91%	56.15%	26.37%	91.34%
	F	58.13%	52.58%	**62.52**%	56.75%	41.08%	**65.83**%	56.15%	39.45%	**65.63**%
2	P	59.52%	58.78%	62.16%	66.07%	64.66%	69.23%	69.64%	70.00%	69.12%
	R	59.52%	84.62%	29.87%	66.07%	82.42%	46.75%	69.64%	76.92%	61.04%
	F	**59.52**%	69.37%	40.35%	**66.07**%	**72.46**%	55.81%	**69.64**%	73.30%	64.83%
3	P	58.93%	58.05%	62.77%	65.08%	63.43%	69.23%	69.64%	69.35%	70.10%
	R	58.93%	87.18%	25.54%	65.08%	83.88%	42.86%	69.64%	78.75%	58.87%
	F	58.93%	69.69%	36.31%	65.08%	72.24%	52.94%	**69.64**%	73.76%	64.00%
4	P	57.94%	57.08%	63.01%	64.68%	62.80%	69.92%	69.25%	68.91%	69.79%
	R	57.94%	90.11%	19.91%	64.68%	85.35%	40.26%	69.25%	78.75%	58.01%
	F	57.94%	**69.89**%	30.26%	64.68%	72.36%	51.10%	69.25%	73.50%	63.36%
5	P	57.54%	56.78%	62.32%	64.09%	62.30%	69.23%	69.64%	68.75%	71.20%
	R	57.54%	90.48%	18.61%	64.09%	85.35%	38.96%	69.64%	80.59%	56.71%
	F	57.54%	69.77%	28.67%	64.09%	72.02%	49.86%	**69.64**%	**74.20**%	63.13%

TABLE 5.6. Spanish; Precision (P), Recall (R) and F-measure (F) for the bootstrapping subjectivity lexicon over 5 iterations and an LSA threshold of 0.5

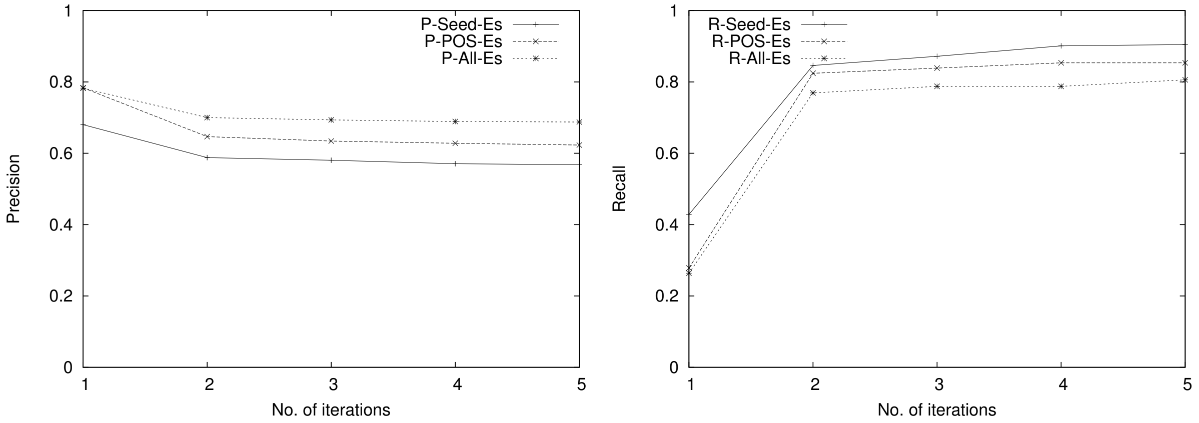
recall level obtained for Romanian when using the *all* variation. This anomaly results from the extremely low coverage of the Romanian subjectivity lexicon extracted by this method (at most 300 entries), and should one have had access to a more elaborate Romanian dictionary, and through it to a reacher set of candidates, I expect the behavior of the recall curves to correct and be consistent



(a) Precision

(b) Recall

FIGURE 5.6. Romanian; Subjectivity precision and recall over five iterations.



(a) Precision

(b) Recall

FIGURE 5.7. Spanish; Subjectivity precision and recall over five iterations.

in the three variations of the method across the two languages. The experiments support the conclusion that if all the bootstrapping variations are able to extract a lexicon of around 1000 entries, then the best results should be obtained by the setup enforcing the strictest similarity (*all*).

Furthermore, Tables 5.5 and 5.6 show that the best overall F-measure and the best subjective F-measure seem to be obtained using the lexicon generated after the fifth iteration, which provides a consistent classification pattern. For Spanish, for example, the highest subjective F-measure obtained in the last iteration is 74.20%, whereas for Romanian is 65.25%.

		Seed			POS			All		
LSA	Eval	Overall	Subj.	Obj.	Overall	Subj.	Obj.	Overall	Subj.	Obj.
0.4	P	62.50%	61.17%	66.41%	62.10%	70.71%	56.54%	60.71%	81.51%	54.29%
	R	62.50%	84.25%	36.80%	62.10%	51.28%	74.89%	60.71%	35.53%	90.48%
	F	62.50%	**70.88**%	47.35%	**62.10**%	**59.45**%	64.43%	**60.71**%	**49.49**%	**67.86**%
0.45	P	64.48%	65.56%	62.87%	58.93%	74.63%	53.24%	56.75%	82.35%	51.55%
	R	64.48%	72.53%	54.98%	58.93%	36.63%	85.28%	56.75%	25.64%	93.51%
	F	**64.48**%	68.87%	58.66%	58.93%	49.14%	**65.56**%	56.75%	39.11%	66.46%
0.5	P	64.29%	68.98%	59.85%	56.94%	76.42%	51.76%	52.38%	82.35%	49.01%
	R	64.29%	61.90%	67.10%	56.94%	29.67%	89.18%	52.38%	15.38%	96.10%
	F	64.29%	65.25%	63.27%	56.94%	42.74%	65.50%	52.38%	25.93%	64.91%
0.55	P	63.49%	75.43%	57.14%	54.37%	78.67%	50.12%	50.99%	86.11%	48.29%
	R	63.49%	48.35%	81.39%	54.37%	21.61%	93.07%	50.99%	11.36%	97.84%
	F	63.49%	58.93%	**67.14**%	54.37%	33.91%	65.15%	50.99%	20.06%	64.66%

TABLE 5.7. Romanian; Precision (P), Recall (R) and F-measure (F) for the 5th bootstrapping iteration for varying LSA scores

		Seed			POS			All		
LSA	Eval	Overall	Subj.	Obj.	Overall	Subj.	Obj.	Overall	Subj.	Obj.
0.4	P	55.56%	55.20%	60.61%	60.12%	58.53%	68.29%	64.09%	61.92%	71.19%
	R	55.56%	95.24%	8.66%	60.12%	90.48%	24.24%	64.09%	87.55%	36.36%
	F	55.56%	69.89%	15.15%	60.12%	71.08%	35.78%	64.09%	**72.53**%	48.14%
0.45	P	67.26%	65.34%	71.71%	56.94%	56.14%	64.58%	62.70%	60.65%	70.48%
	R	67.26%	84.25%	47.19%	56.94%	93.77%	13.42%	62.70%	88.64%	32.03%
	F	**67.26**%	**73.60**%	**56.92**%	56.94%	70.23%	22.22%	62.70%	72.02%	44.05%
0.5	P	64.09%	62.30%	69.23%	69.64%	68.75%	71.20%	57.54%	56.78%	62.32%
	R	64.09%	85.35%	38.96%	69.64%	80.59%	56.71%	57.54%	90.48%	18.61%
	F	64.09%	72.02%	49.86%	**69.64**%	**74.20**%	**63.13**%	57.54%	69.77%	28.67%
0.55	P	61.11%	59.85%	65.49%	64.48%	63.35%	67.11%	68.06%	74.14%	62.87%
	R	61.11%	85.71%	32.03%	64.48%	81.68%	44.16%	68.06%	63.00%	74.03%
	F	61.11%	70.48%	43.02%	64.48%	71.36%	53.26%	**68.06**%	68.12%	**67.99**%

TABLE 5.8. Spanish; Precision (P), Recall (R) and F-measure (F) for the 5th bootstrapping iteration for varying LSA scores

To examine the effect of the number of bootstrapping iterations and the value of the LSA similarity threshold on the classifier, Tables 5.5 and 5.6 display the measures obtained through five bootstrapping iterations at an LSA threshold of 0.50, while Tables 5.7 and 5.8 focus on the fifth iteration tested over an LSA similarity of 0.40, 0.45, 0.50, and 0.55. As expected, the overall F-measure is directly proportional to the LSA similarity score until the threshold becomes too restrictive, explicitly limiting the number of entries in the subjectivity lexicon. The variation of the LSA score produces at most 6% fluctuation in both overall and subjective F-measures for both Spanish and Romanian (with the exception of the *POS* and *all* variations). The results indicate that

	Romanian			Spanish		
Eval	Overall	Subj.	Obj.	Overall	Subj.	Obj.
Most Frequent Sense						
P	55.95%	65.84%	51.31%	58.53%	66.00%	53.62%
R	55.95%	38.83%	76.19%	58.53%	48.35%	70.56%
F	55.95%	**48.85**%	61.32%	58.53%	55.81%	60.93%
Part-of-speech Based Sense						
P	56.75%	67.97%	51.85%	59.72%	66.99%	54.70%
R	56.75%	38.10%	78.79%	59.72%	50.55%	70.56%
F	**56.75**%	48.83%	**62.54**%	**59.72**%	**57.62**%	**61.63**%

TABLE 5.9. Precision (P), recall (R) and F-measure (F) for the automatic translation of the subjectivity lexicon of strength weak and strong

a LSA threshold of 0.5 achieves a consistent subjective F-measure as long as it allows sufficient candidates to aggregate in the subjectivity lexicon.

The results are compared with those obtained by automatically translating a subjectivity lexicon, as described in Section 5.1. In that method, a subjectivity lexicon is automatically obtained through the translation of the *strong* and *weak* entries composing the English subjectivity lexicon available in OpinionFinder. The translation focuses on automatically acquiring *the most frequent sense*, regardless of the part-of-speech information adjoining the original English entry. Two lexicons are obtained using this method: a Romanian lexicon consisting of 1,580 entries, and a Spanish lexicon of 2,009 entries, respectively. These are automatically evaluated using the same rule-based classifier and the same gold-standard data-set as used in testing the bootstrapping method. The results are shown in the top part of Table 5.9.

I also conduct an additional experiment (*part-of-speech based sense*) to test whether the automatic translation of the lexicon may perform better if the part-of-speech annotation from the English subjectivity lexicon is utilized to disambiguate among candidate translations in the target language. Starting with the subjectivity lexicon annotated with the corrected part-of-speech tag (explained in Section 5.1), I then use the same Ectaco dictionary for performing lexicon translations into Romanian and Spanish. If the online dictionary offers a translation candidate for the part-of-speech projected from English, then this is accepted as the first entry among the translations. Otherwise, the English entry is disregarded. The resulting lexicons in both Romanian and Spanish are about 50 words shorter (Romanian: 1509 entries, Spanish: 1953 entries) when compared to the more generic first sense translation counterparts. Despite the additional effort entailed in correcting the part-of-speech and enforcing it in a bilingual dictionary, the results show only a marginal 1% improvement in F-measure for both languages (see the bottom part of Table 5.9).

In order to asses whether the best option to extract a subjectivity lexicon in a target language is by direct translation of a source language subjectivity lexicon, or by growing the target language lexicon from a small number of seeds, let us compare Tables 5.5 and 5.6, showing the bootstrapping results, with Table 5.9, which shows the lexicon translation results. The overall F-measure obtained for Romanian after the fifth iteration, a LSA threshold of 0.5 and the *seed* variation is 64.29%, being higher by 8.34% or 7.54% when compared to the *most frequent sense* or the *part-of-speech based sense* overall F-measure, respectively. The results for Spanish are even more compelling, as the best bootstrapping results achieved under the *POS* variation is 69.64% F-measure, while the *most frequent sense* and the *part-of-speech based sense* reach 58.53% and 59.72%. It is also important to note that the bilingual dictionary extracts more candidates than the bootstrapping process is able to, yet they appear less frequently in the target language, therefore hampering the recall of the rule-based classifier, and undermining a high subjectivity F-measure. These aspects support the conclusion that a more reliable subjectivity lexicon can be extracted directly in a target language, instead of requiring a sizeable subjectivity lexicon in a source language and a bridging bilingual dictionary.

CHAPTER 6

SENTENCE LEVEL MULTILINGUAL SUBJECTIVITY

This chapter explores the potential offered by sentence level subjectivity annotations (whether manual or automatic) to participate in a machine learning framework. Section 6.1 will focus on leveraging monolingual projections individually, while Section 6.2 will seek to gauge the cumulative effect that multilingual data has on sentence level subjectivity.

Parts of this chapter were previously published in [10, 6, 8].

6.1. Monolingual Learning

This section will focus on formulating an answer to the following questions:

Question 1: If an English corpus manually annotated for subjectivity is available, can we use machine translation to generate a subjectivity-annotated corpus in the target language and train a subjectivity classifier in the target language?

Question 2: Assuming the availability of a tool for automatic subjectivity analysis in English, can we generate a corpus annotated for subjectivity in the target language by using automatic subjectivity annotations of English text or automatic annotations of automatic parallel text generated through machine translation; can these automatically generated corpora be used to train a subjectivity classifier in the target language?

In order to perform a comprehensive investigation, I propose four experiments as described below. The first scenario, based on a corpus manually annotated for subjectivity, is exemplified by the first experiment. The second scenario, based on a corpus automatically annotated with a tool for subjectivity analysis, is subsequently divided into three experiments depending on the type of parallel text (manual or automatic), and the direction of the translation. Given a source and a target language, the choice of the best scenario to be used is to be made depending on the resources available for that source and target language.

In all four experiments, English is used as a source (or donor) language, given that it has both a corpus manually annotated for subjectivity (MPQA [80]) and a tool for subjectivity analysis (OpinionFinder [77]).

6.1.1. Manually Annotated Corpora

6.1.1.1. Experiment One: Machine Translation of Manually Annotated Corpora

Question 1: If an English corpus manually annotated for subjectivity is available, can we use machine translation to generate a subjectivity-annotated corpus in the target language and train a subjectivity classifier in the target language?

In this experiment, we start from a corpus in the source language which is manually annotated for subjectivity. The data is transferred into the target language through the intercession of a machine translation engine, and then it is augmented with subjectivity labels projected from the source language annotations.

The experiment is illustrated in Figure 6.1.

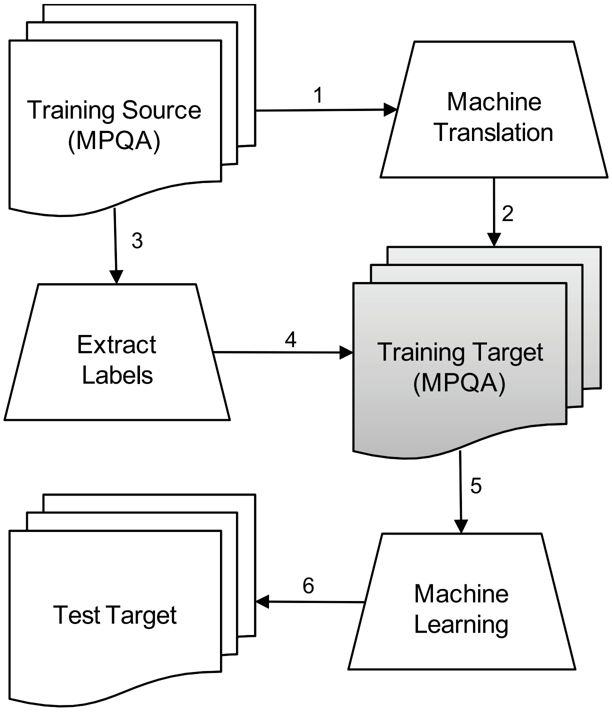


FIGURE 6.1. Experiment 1: machine translation of manually annotated training data from source language into target language.

We use the MPQA corpus (see Section 3.1.2). After the automatic translation of the corpus and the projection of the annotations, we obtain a large corpus of 9,700 subjectivity-annotated sentences in the target language, which can be used to train a subjectivity classifier.

6.1.2. Automatically Annotated Corpora

Question 2: Assuming the availability of a tool for automatic subjectivity analysis in English, can we generate a corpus annotated for subjectivity in the target language by using automatic subjectivity annotations of English text or automatic annotations of automatic parallel text generated through machine translation; can these automatically generated corpora be used to train a subjectivity classifier in the target language?

A subset of the English SemCor corpus [44] consisting of 107 documents with roughly 11,000 sentences (which do not include the gold-standard data set described in Section 5.3) serves as a raw corpus. This is a balanced corpus covering a number of topics in sports, politics, fashion, education, and others. The reason for working with this collection is the fact that we also have a manual translation of the SemCor documents from English into one of the target languages used in the experiments (Romanian), which enables comparative evaluations of different scenarios (see Section 6.1.3). This text (Sections 6.1.2.1 and 6.1.2.2) or a machine translated version of it into English (Section 6.1.2.3) is processed with the aid of the high-coverage classifier embedded in the OpinionFinder tool (see Section 3.1.3), resulting in automatic sentence level subjectivity annotations.

6.1.2.1. Experiment Two: Manually Translated Parallel Text

The second experiment assumes that we have access to a tool for subjectivity annotation developed for the source language, and to a manually constructed parallel text in both the source and target language. This tool is able to generate automatic subjectivity annotations for the source data set, which can be projected via the parallel text into the target language. This scenario results in an automatically annotated subjectivity data set in the target language, which can be used to train an automatic classifier in that language. This experiment is exemplified in Figure 6.2.

Since this experiment uses manually translated parallel text, both the subjectivity annotation tool in the source language and the training of the classifier in the target language are performed on correct human generated text. Therefore this setup should provide better results when compared to the machine translation experiments.

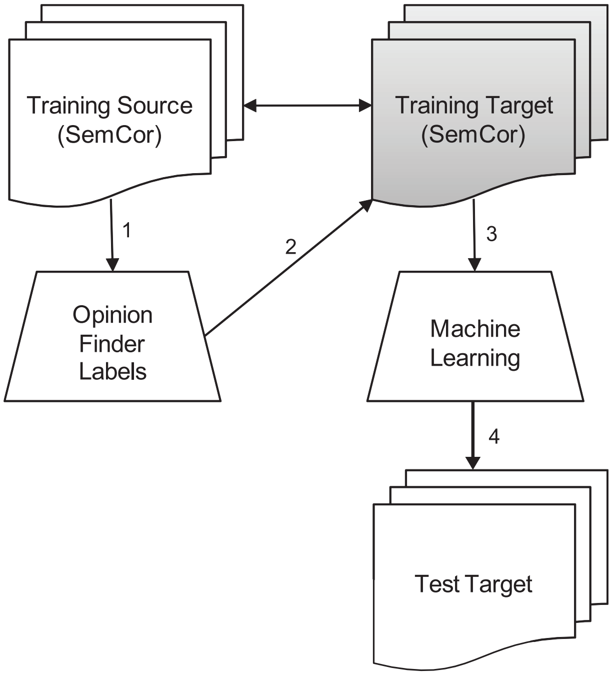


FIGURE 6.2. Experiment 2: manually translated parallel text.

6.1.2.2. Experiment Three: Machine Translation of Source Language Training Data

The third experiment covers the scenario where the only resources available are a tool for subjectivity annotation in the source language and a collection of raw texts, also in the source language. The source language text is automatically annotated for subjectivity and then translated into the target language. In this way, a subjectivity annotated corpus can be produced and then ultimately used to train a subjectivity annotation tool for the target language. Figure 6.3 illustrates this experiment.

Note that in this experiment the annotation of subjectivity is carried out on the original source language text, and thus expected to be more accurate than if it were applied on automatically translated text. However, the training data in the target language is produced by automatic translation, and thus likely to contain errors.

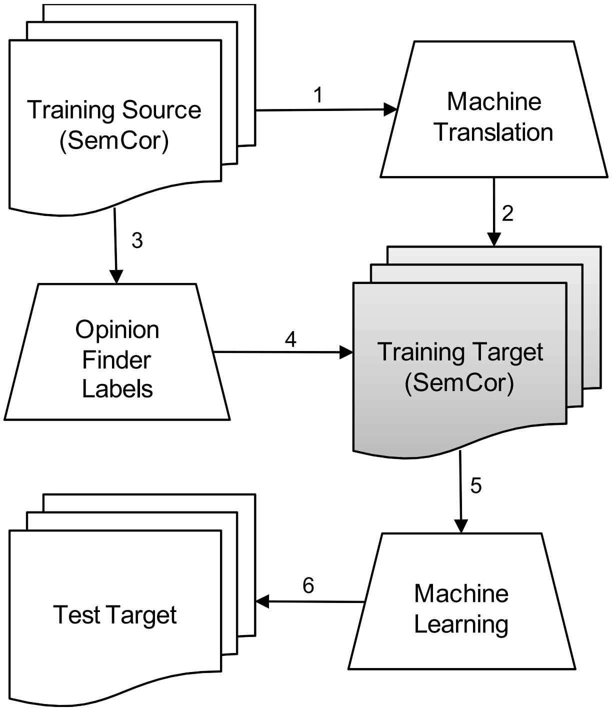


FIGURE 6.3. Experiment 3: machine translation of raw training data from source language into target language.

6.1.2.3. Experiment Four: Machine Translation of Target Language Training Data

The fourth experiment is similar to the third one, except that the translation direction is reversed. Raw text that is available in the target language is translated into the source language; the automatically translated source language text is then annotated with a subjectivity annotation tool. After the annotation, the labels are projected back into the target language, and the resulting annotated corpus is used to train a subjectivity classifier. Figure 6.4 illustrates this experiment.

In this experiment, the subjectivity annotations are carried out on automatically generated source text, and thus expected to be less accurate. However, since the training data was originally written in the target language, it is free of translation errors, and thus training carried out on this data should be more robust.

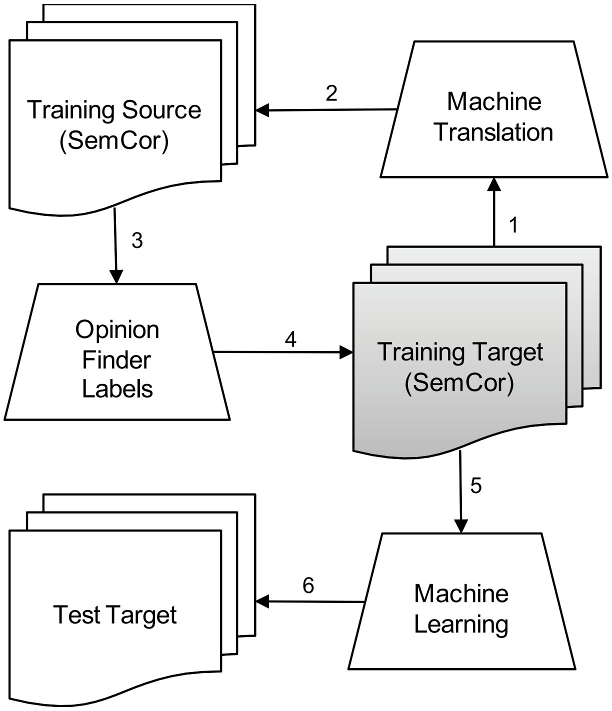


FIGURE 6.4. Experiment four: machine translation of raw training data from target language into source language.

6.1.2.4. Upper bound: Machine Translation of Target Language Test Data

For comparison purposes, I also propose an experiment which plays the role of an upper bound on the machine translation methods proposed in Sections 6.1.2.2 and 6.1.2.3. It involves the automatic translation of the test data from the target language into the source language. The source language text is then annotated for subjectivity using OpinionFinder, followed by the projection of the resulting labels back into the target language.

Unlike the previous experiments, this setup only generates subjectivity-annotated *resources*, and it is not used to build and evaluate a standalone subjectivity analysis *tool* for the target language. Further training of a machine learning algorithm, as in experiments three and four, is required in order to build a subjectivity analysis tool. Thus, this study is an evaluation of the *resources* generated in the target language, thus representing an upper bound on the performance of any machine learning algorithm that would be trained on these resources. Figure 6.5 illustrates this experiment.

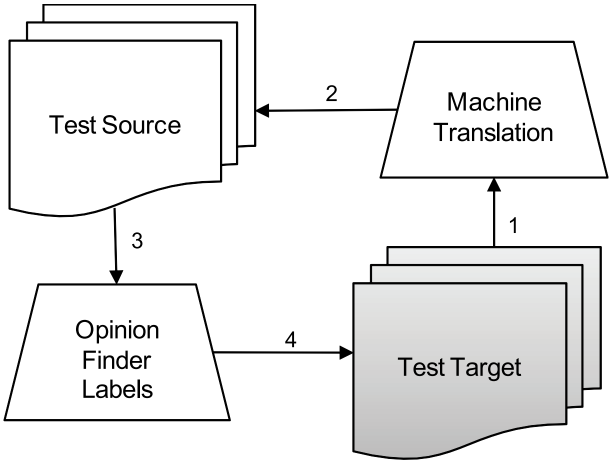


FIGURE 6.5. Upper bound: machine translation of test data from target language into source language.

6.1.3. Evaluation and Results

The evaluations are carried out on Romanian and Spanish (where data availability allows) and the performance of each of the four experiments proposed in this section is evaluated using the same gold-standard described in Section 5.3. To evaluate the methods, a training corpus annotated for subjectivity is generated based on projections from the source language, for both Romanian and Spanish. The document-label pairs are passed to a machine learner under the hypothesis that these labels are the accurate annotations for the target sentences. The underlying assumption is that any possible translation or annotation errors in the training data will be eventually voted out during the training stage of the classifiers. For learning, the framework uses a state-of-the-art algorithm, namely Support Vector Machines.

Support Vector Machines (SVM) [72, 27] is a machine learning approach based on decision planes. The algorithm seeks to render the optimal hyper-plane that separates the set of points associated with different class labels resulting in a maximum-margin. The unlabeled examples are then classified by deciding on which side of the hyper-surface they reside. Such an algorithm is most advantageous with a noisy training data such as the one provided in the current scenario. The evaluations use the implementation available in the AI::Categorizer module mentioned above with a linear kernel, since it was proved to be as powerful as other kernels in text classification experiments [84].

A feature selection algorithm is also applied, keeping only the top 50% discriminating features, according to a $tf.idf$ weighting scheme using raw term frequencies normalized by the length of the document vector [54].

For Romanian, the automatic translation of the MPQA and of the SemCor corpus was performed using Language Weaver,[1] a commercial statistical machine translation software. To generate the training data required for the equivalent experiments on Spanish, both the MPQA corpus and the SemCor corpus are translated using the Google Translation service,[2] a publicly available machine translation engine also based on statistical machine translation. Since a manual translation into Spanish of the SemCor training data was not available, experiments two and four were not be replicated in this language. Although any Spanish text could have been selected to carry out a similar experiment, due to the fact that the dataset would have been different, the results would not have been directly comparable. The resulting text in the target languages was post-processed by removing diacritics, stopwords and numbers.

The results obtained by running the four experiments on Romanian and Spanish are shown in Table 6.1. The baseline on this data set is 54.16%, represented by the percentage of sentences in the corpus that are subjective, and the upper bound (UB) is 71.83%, which is the accuracy obtained under the scenario where the test data is translated into the source language and then annotated using the high-coverage OpinionFinder tool. All the results are statistically significant at $p < 0.05$ when compared to a random output following the class distribution present in the goldstandard.

For Romanian, the first experiment, involving the automatic translation of the MPQA corpus enhanced with manual annotations for subjectivity at sentence level, does not seem to perform well when compared to the experiments in which automatic subjectivity classification is used on either manual or automatically generated text. This could imply that a classifier cannot be so easily trained on the cues that humans use to express subjectivity, especially when they are not overtly expressed in the sentence and thus can be lost in the translation. Instead, the automatic annotations produced with a rule-based tool (OpinionFinder), relying on overt mentions of words in a

[1] http://www.languageweaver.com/

[2] http://www.google.com/translate_t

			Romanian			Spanish		
Classifier	Exp	Eval	Overall	Subj.	Obj.	Overall	Subj.	Obj.
SVM	E1	P	66.07%	59.00%	70.73%	68.85%	67.90%	70.56%
		R	66.07%	91.21%	25.11%	68.85%	80.59%	54.98%
		F	66.07%	71.65%	37.06%	**68.85**%	**73.70**%	61.80%
	E2	P	69.64%	69.35%	70.10%			
		R	69.64%	78.75%	58.87%			
		F	69.64%	73.76%	64.00%			
	E3	P	69.44%	67.76%	72.78%	63.89%	73.82%	57.83%
		R	69.44%	83.15%	53.25%	63.89%	51.65%	78.35%
		F	**69.44**%	**74.67**%	61.50%	63.89%	60.78%	**66.54**%
	E4	P	67.86%	76.06%	61.86%			
		R	67.86%	59.34%	77.92%			
		F	67.86%	66.67%	**68.97**%			
OpinionFinder	UB	P	71.83%	71.91%	71.71%	73.41%	73.88%	72.77%
		R	71.83%	78.75%	63.64%	73.41%	78.75%	67.10%
		F	71.83%	75.17%	67.43%	73.41%	76.24%	69.82%

TABLE 6.1. Precision (P), Recall (R) and F-measure (F) for Romanian and Spanish experiments;
Manual subjectivity annotations: E1 - source to target language machine translation;
Automatic subjectivity annotations: E2 - parallel text, E3 - source to target language machine translation, E4 - target to source language machine translation.

subjectivity lexicon, seem to be more robust to translation, further resulting in better classification results. To exemplify, let us consider the following subjective sentence from the MPQA corpus, which does not include overt clues of subjectivity, but was annotated as subjective by the human judges because of the structure of the sentence:

> It is the Palestinians that are calling for the implementation of the agreements, understandings, and recommendations pertaining to the Palestinian-Israeli conflict. (6)

A unigram classifier would not be able to determine that the author of this sentence is using word topology to express his opinion about the conflict. The writer is able to state and emphasize his perspective even though it is not overtly expressed in the sentence by any given word. A learning algorithm will not find the above sentence very different from *Palestinians are calling for the implementation of the agreements, understandings, and recommendations pertaining to the Palestinian-Israeli conflict.*, especially once stopwords are removed.

The results of the two experiments carried out on Spanish are shown in Table 6.1. Interestingly, the F-measure is higher for the first experiment involving the machine translated version of a corpora manually labeled for subjectivity, than its counterpart in Romanian. This probably happens because Spanish is one of the six official United Nations languages, thus having larger amounts of parallel texts available to train the machine translation system, which implies that a better quality translation can be achieved as compared to the one available for Romanian. Since the Spanish automatic translation seems to be closer to a human-quality translation, it is not surprising that this time the first experiment is able to generate a more accurate training corpus as compared to the third experiment, surpassing the overall F-measure calculated for the third experiment by 4.96%. The MPQA corpus, since it is manually annotated and of better quality, has a higher chance of generating a more reliable data set in the target language. Unlike the results obtained for Romanian in experiments three and four, in Spanish, the classifier is not able to distinguish as well the subjective cases, reaching a subjectivity F-measure of only 60.78%, and thus penalizing the overall F-measure. As in the experiments on Romanian, when performing automatic translation of the test data, the best results attain an F-measure of 73.41%, which represents the upper bound on our proposed experiments.

Among the approaches proposed in this section, experiments three and four are closest to the experiment based on parallel text (second experiment). By using machine translation, from English into Romanian or Spanish (experiment three) or Romanian into English (experiment four), and annotating this dataset with the high-coverage OpinionFinder classifier using an SVM learning algorithm, we obtain an F-measure of 63.89% / 69.44% (experiment three), and 67.86% respectively (experiment four). This implies that using a parallel corpus does not produce significantly better results when compared to automatically generated translations, especially when the training set was automatically annotated. This finding further suggests that machine translation is a viable alternative to devising subjectivity classification in a target language leveraged on the tools existent in a source language. Despite the loss in readability quality of the resulting automatic text, its subjectivity content seems to be mostly rendered correctly. The statistical machine translation

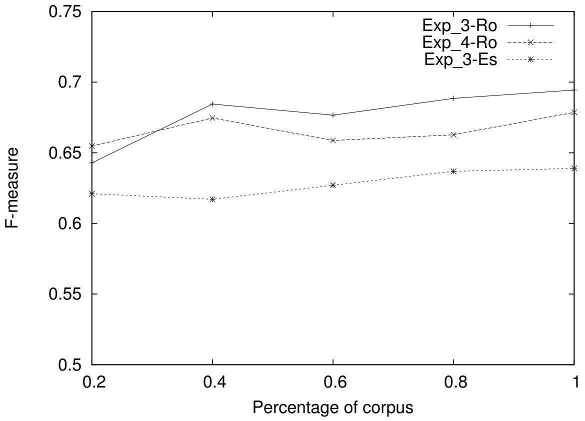


FIGURE 6.6. Machine learning F-measure over an incrementally larger training set in Romanian (Ro) and Spanish (Es); automatic subjectivity annotations; Experiment three (Exp_3) - source to target language machine translation; Experiment four (Exp_4) - target to source language machine translation.

engines are better equipped to transfer the subjective sense of a word since they disambiguate it based on n-gram language models.

To illustrate this argument, consider the following example in English and its translation into Romanian and Spanish using Google Translate:

En: “I *boil* with anger.”

Ro: “Am *fiert* cu furie.”

Es: “Me *hierve* con rabia.”

If the direct translation experiments described in Section 5.1 translate the verb *boil* into noun *Ro: fierbere* (as in *He brought a kettle of water to a boil.*[3]) and noun *Es: furúnculo* (with the meaning of *En: furuncle*), the machine translation correctly identifies that *boil* is used in its verb sense. Even though the preposition (*cu*) in Romanian is not the one a native would employ, the translation of the sentence ensures that its subjective meaning, and components, are accurately rendered into the target language.

Finally, we also wanted to explore the impact that the corpus size may have on the accuracy of the classifiers. We re-ran experiments three and four with 20% corpus size increments at a time (Figure 6.6). It is interesting to note that a corpus of approximately 6000 sentences is able to

[3] example obtained from http://dictionary.reference.com/

achieve a high enough F-measure (around 66% for both experiments in Romanian and 63% for the Spanish experiment) to be considered viable for training a subjectivity classifier.

6.2. Multilingual Learning

The second part of this chapter builds upon the conclusions attained from using monolingual learning (see Section 6.1 in a sentence level subjectivity context. Additional monolingual experiments are carried out to further strengthen the prior observations, while also delving deeper into using languages jointly to perform subjectivity classification.

These are the questions I aim to answer:

Question 1: Can sentence-level subjectivity be *reliably* predicted in languages other than English, by leveraging on a manually annotated English dataset?

Question 2: Can English subjectivity classification be improved by expanding the feature space through the use of multilingual data? Similarly, can the classifiers in the other target languages experience an improvement as well?

Question 3: Could the multilingual subjectivity space be used to train a high-precision subjectivity classifier that could ultimately generate subjectivity datasets in the target languages?

6.2.1. Multilingual Datasets

Building upon the method proposed earlier to generate parallel annotated data in other languages, I will reiterate experiment one (see Section 6.1.1.1) involving the automatic translation of a manually annotated dataset and experiment with five languages other than English (*En*), namely Arabic (*Ar*), French (*Fr*), German (*De*), Romanian (*Ro*) and Spanish (*Es*). The choice of languages is motivated by several reasons. First, I was interested in using languages that are highly lexicalized and have clear word delimitations. Second, I wanted to cover languages that are similar to English as well as languages with a completely different etymology. Consideration was given to include Asian languages, such as Chinese or Japanese, but the fact that their script without word-segmentation preprocessing does not directly map to words was a deterrent. Finally, another limitation on the choice of languages is the need for a publicly available machine translation system between the source language and each of the target languages.

The donor corpus in this case remains the English Multi-Perspective Question Answering (MPQA) corpus described in Section 3.1.2 and its sentence-level annotations. From the approximately 9700 sentences in this corpus, 55% of them are labeled as subjective, while the rest are objective. Therefore, 55% represents the majority baseline on this corpus.

A subjectivity annotated corpus is constructed for each of the five languages by using machine translation to transfer the source language data into the target language[4]. The original sentence level English subjectivity labeling are then projected onto the target data.

For all languages, other than Romanian, I use the Google Translate service,[5] a publicly available machine translation engine based on statistical models. The reason Romanian is not included in this group is that, at the time the first experiments were performed, Google Translate did not provide a translation service for this language. Therefore, as mentioned in the previous section, the Romanian translation is obtained by using an alternative commercially available statistical translation system called LanguageWeaver,[6], and which the company kindly provided access to for research purposes.

Despite the fact that the current experiments use the same MPQA translation (for Spanish and Romanian) as the ones from the previous section, the results are not directly comparable. Here the evaluations are performed through cross validation on the entire MPQA set using the labels proposed for English by [77]; in Section 6.1, the test set consisted of a separate gold standard of 504 sentences extracted from SemCor (see Section5.3), that were originally annotated in Romanian.

Given the specifics of each language, several preprocessing techniques are employed. In the case of Romanian, French, English, German and Spanish, all diacritics, numbers and punctuation marks except - and ' are removed. The exceptions are motivated by the fact that they may mark contractions, such as En: *it's* or Ro: *s-ar* (*may be*), and the component words may not resolve to the correct forms. For Arabic, although it has a different encoding, I wanted to make sure to treat it in a way similar to the languages with a Roman alphabet. The text was thus preprocessed

[4] The raw corpora in the five target languages are available for download at `http://lit.csci.unt.edu/index.php/Downloads`.

[5] http://www.google.com/translate_t

[6] http://www.languageweaver.com/

using a library[7] that maps Arabic script to a space of Roman-alphabet letters supplemented with punctuation marks so that they can allow for the additional dimensionality.

Let us consider the letter “s” which has three versions in Arabic:

Arabic letter	Unicode name	Transliteration
س	seen	s
ش	sheen	ˆs
ص	sad	.s

Note the transliteration composed of a pair: an optional symbol succeeded by an English alphabet letter.

Once the corpora are preprocessed, each sentence is defined by six views: one in the original source language (English), and five obtained through automatic translation in each of the target languages. Multiple datasets that cover all possible combinations of six languages taken one through six (a total of 63 combinations) are generated. These datasets feature a vector for each sentence present in MPQA (approximately 9700). The vector contains only unigram features in one language for a monolingual dataset. For a multilingual dataset, the vector represents a cumulation of monolingual unigram features extracted from each view of the sentence. For example, one of the combinations of six taken three is Arabic-German-English. For this combination, the vector is composed of unigram features extracted from each of the Arabic, German and English translations of the sentence.

The evaluations are performed using ten-fold cross validation upon training Naïve Bayes classifiers with feature selection on each dataset combination. The top 20% of the features present in the training data based on document frequency are retained. For datasets resulting from combinations of all languages taken one, the classifiers are monolingual classifiers. All other classifiers are multilingual, and their feature space increases with each additional language added. Expanding the feature set by encompassing a group of languages enables us to provide an answer to two problems that can appear due to data sparseness. First, enough training data may not be available in the monolingual corpus alone in order to correctly infer labeling based on statistical measures.

[7] Lingua::AR::Word PERL library.

Second, features appearing in the monolingual test set may not be present in the training set and therefore their information cannot be used to generate a correct classification.

Both of these problems are further explained through the examples below, with the simplifying assumption that the words in italics are the only potential carriers of subjective content, and that, without them, their surrounding contexts would be objective. Therefore, their association with an either objective or subjective meaning imparts to the entire segment the same labeling upon classification.

To explore the first sparseness problem, let us consider the following two examples extracted from the English version of the MPQA dataset, followed by their machine translations in German:

En 1: “rights group Amnesty International said it was *concerned* about the high risk of violence in the aftermath”

En 2: “official said that US diplomats to countries *concerned* are authorized to explain to these countries”

De 1: “Amnesty International sagte, es sei *besorgt* über das hohe Risiko von Gewalt in der Folgezeit”

De 2: “Beamte sagte, dass US-Diplomaten *betroffenen* Länder berechtigt sind, diese Länder zu erklären”

Focusing our discussion on the word *concerned*, the first example showcases its usage in a subjective sense, while in the second it carries an objective meaning (as it refers to a group of countries exhibiting a particular feature defined earlier on in the context). The words in italics in the German contexts represent the translations of *concerned* into German, which are functionally different as they are shaped by their surrounding context. By training a classifier on the English examples alone, under the data sparseness paradigm, the machine learning model may not differentiate between the word’s objective and subjective uses when predicting a label for the entire

Lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF	MAcc
En	74.01%	**83.64**%	**78.53**%	**75.89**%	63.68%	**69.25**%	**74.95**%	**73.66**%	**73.89**%	**74.72**%
Ro	73.50%	82.06%	77.54%	74.08%	63.40%	68.33%	73.79%	72.73%	72.94%	73.72%
Es	**74.02**%	82.84%	78.19%	75.11%	**64.05**%	69.14%	74.57%	73.44%	73.66%	74.44%
Fr	73.83%	83.03%	78.16%	75.19%	63.61%	68.92%	74.51%	73.32%	73.54%	74.35%
De	73.26%	83.49%	78.04%	75.32%	62.30%	68.19%	74.29%	72.90%	73.12%	74.02%
Ar	71.98%	81.47%	76.43%	72.62%	60.78%	66.17%	72.30%	71.13%	71.30%	72.22%

TABLE 6.2. Naïve Bayes learners trained on six individual languages.

sentence. However, appending the German translation to the examples generates additional dimensions for this model and allows the classifier to potentially distinguish between the senses and provide the correct sentence label.

For the second problem, let us consider two other examples from the English MPQA and their respective translations into Romanian:

En 3: "could secure concessions on Taiwan in return for *supporting* Bush on issues such as anti-terrorism and"

En 4: "to the potential for change from within America. *Supporting* our schools and community centres is a good"

Ro 3: "ar putea asigura concesii cu privire la Taiwan, în schimb pentru *susţinerea* lui Bush pe probleme cum ar fi anti-terorismului şi"

Ro 4: "la potenţialul de schimbare din interiorul Americii. *Sprijinirea* şcolile noastre şi centre de comunitate este un bun"

In this case, *supporting* is used in both English examples in senses that are both subjective; the word is, however, translated into Romanian through two synonyms, namely *susţinerea* and *sprijinirea*. Let us assume that sufficient training examples are available to strengthen a link between *supporting* and *susţinerea*, and the classifier is presented with a context containing *sprijinirea*, unseen in the training data. A multilingual classifier may be able to predict a label for the context using the co-occurrence metrics based on *supporting* and extrapolate a label when the context contains both the English word and its translation into Romanian as *sprijinirea*. For a monolingual classifier, such an inference is not possible, and the feature is discarded. Therefore a multi-lingual classifier model may gain additional strength from co-occurring words across languages.

No lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF
1	73.43%	82.76%	77.82%	74.70%	62.97%	68.33%	74.07%	72.86%	73.08%
2	74.59%	83.14%	78.63%	75.70%	64.97%	69.92%	75.15%	74.05%	74.28%
3	75.04%	83.27%	78.94%	76.06%	65.75%	70.53%	75.55%	74.51%	74.74%
4	75.26%	83.36%	79.10%	76.26%	66.10%	70.82%	75.76%	74.73%	74.96%
5	75.38%	83.45%	79.21%	76.41%	66.29%	70.99%	75.90%	74.87%	75.10%
6	**75.43%**	**83.66%**	**79.33%**	**76.64%**	**66.30%**	**71.10%**	**76.04%**	**74.98%**	**75.21%**

TABLE 6.3. Average measures for a particular number of languages in a combination (from one through six) for Naïve Bayes classifiers using a multilingual space.

6.2.2. Evaluation and Results

In this section I revisit the three questions formulated originally.

6.2.2.1. Question 1

Question 1: Can sentence-level subjectivity be *reliably* predicted in languages other than English, by leveraging on a manually annotated English dataset?

In Section 6.1, several methods for porting subjectivity annotated data from a source language (English) to a target language (Romanian and Spanish) were explored. Here, the focus lies on the transfer of manually annotated corpora through the usage of machine translation by projecting the original sentence level annotations onto the generated parallel text in the target language. The aim is not to improve on that method, but rather to verify that the results are reliable across a number of languages. Therefore, this experiment is conducted in several additional languages, namely French, German and Arabic, and the results are compared with those obtained for Spanish and Romanian.

Table 6.2 shows the results obtained using Naïve Bayes classifiers trained in each language individually, with a macro accuracy ranging from 71.30% (for Arabic) to 73.89% (for English).[8] As expected, the English machine learner outperforms those trained on other languages, as the original language of the annotations is English. However, it is worth noting that all measures do not deviate by more than 3.27%, implying that classifiers built using this technique exhibit a consistent behavior across languages.

[8] Note that the experiments conducted in Section 6.1 were made on a different test set, and thus the results are not directly comparable.

6.2.2.2. Question 2

Question 2: Can English subjectivity classification be improved by expanding the feature space through the use of multilingual data? Similarly, can the classifiers in the other target languages experience an improvement as well?

Let us turn towards investigating the impact on subjectivity classification of an expanded feature space through the inclusion of multilingual data. In order to methodically assess classifier behavior, multiple datasets containing all possible combinations of one through six languages are generated, as described in Section 6.2.1. Naïve Bayes learners are trained on the multilingual data and the results are averaged per each group comprised of a particular number of languages. For example, one language comprises the six individual classifiers described in Section 6.2.2.1; for a group of three languages, the average is calculated over 20 possible combinations; and so on.

Table 6.3 shows the results of this experiment. We can see that the overall F-measure increases from 73.08% – which is the average over one language – to 75.21% when all languages are taken into consideration (8.6% error reduction). The statistical significance of these results is measured by considering on one side the predictions made by the best performing classifier for one language (i.e., English), and on the other side the predictions made by the classifier trained on the multilingual space composed of all six languages. Using a paired t-test, the improvement was found to be significant at $p = 0.001$. It is worth mentioning that both the subjective and the objective precision measures increase to 75% when more than 3 languages are considered, while the overall recall level stays constant at 74%.

To verify that the improvement is indeed caused by the addition of multilingual features, and it is not a characteristic of the classifier, two other classifiers were tested, namely KNN and Rocchio. Figure 6.7 shows the average macro-accuracies obtained with these classifiers. For all the classifiers, the accuracies of the multilingual combinations exhibit an increasing trend, as a larger number of languages is used to predict the subjectivity annotations. The Naïve Bayes algorithm has the best performance, and a relative error rate reduction in accuracy of 8.25% for the grouping formed of six languages versus one, while KNN and Rocchio exhibit an error rate reduction of 5.82% and 9.45%, respectively. All of these reductions are statistically significant.

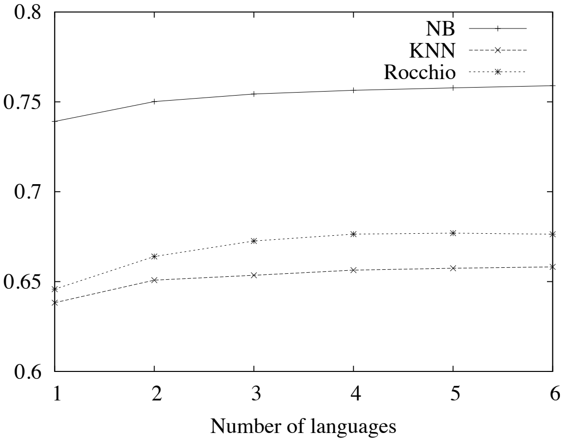


FIGURE 6.7. Average macro-accuracy per group of languages (combinations of 6 taken one through six).

In order to assess how the proposed multilingual expansion improves on the individual language classifiers, one language is selected at a time to be the reference, and then the average accuracies of the Naïve Bayes learner across all the language groupings (from one through six) that contain the language are computed. The results from this experiment are illustrated in Figure 6.8. The baseline in this case is represented by the accuracy obtained with a classifier trained on only one language (this corresponds to 1 on the X-axis). As more languages are added to the feature space, we notice a steady improvement in performance. When the language of reference is Arabic, we obtain an error reduction of 15.27%; 9.04% for Romanian; 7.80% for German; 6.44% for French; 6.06% for Spanish; and 4.90 % for English. Even if the improvements seem minor, they are consistent, and the use of a multilingual feature set enables every language to reach a higher accuracy than individually attainable.

In terms of the best classifiers obtained for each grouping of one through six, English provides the best accuracy among individual classifiers (74.71%). When considering all possible combinations of six classifiers taken two, German and Spanish provide the best results, at 75.67%. Upon considering an additional language to the mix, the addition of Romanian to the German-Spanish classifier further improves the accuracy to 76.06%. Next, the addition of Arabic results in the best performing overall classifier, with an accuracy of 76.22%. Upon adding supplemental languages, such as English or French, no further improvements are obtained. This may be the case

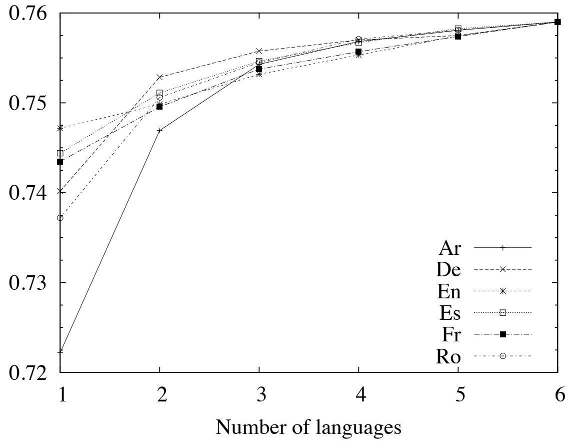


FIGURE 6.8. Average macro-accuracy progression relative to a given language.

because German and Spanish are able to expand the dimensionality conferred by English alone, while at the same time generating a more orthogonal space. Incrementally, Romanian and Arabic are able to provide high quality features for the classification task. This behavior suggests that languages that are somewhat further apart are more useful for multilingual subjectivity classification than intermediary languages.

6.2.2.3. Question 3

Question 3: Could the multilingual subjectivity space be used to train a high-precision subjectivity classifier that could ultimately generate subjectivity datasets in the target languages?

Since the inclusion of multilingual information improves the performance of subjectivity classifiers for all the languages involved, I further explore how the classifiers' predictions can be combined in order to generate high-precision subjectivity annotations. As shown in previous work, a high-precision classifier can be used to automatically generate subjectivity annotated data [52]. Additionally, the data annotated with a high-precision classifier can be used as a seed for bootstrapping methods, to further enrich each language individually.

No lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF
1	73.43%	82.76%	77.82%	74.70%	62.97%	68.33%	74.07%	72.86%	73.08%
2	76.88%	76.39%	76.63%	80.17%	54.35%	64.76%	78.53%	65.37%	70.69%
3	78.56%	72.42%	75.36%	82.58%	49.69%	62.02%	80.57%	61.05%	68.69%
4	79.61%	69.50%	74.21%	84.07%	46.54%	59.89%	81.84%	58.02%	67.05%
5	80.36%	67.17%	73.17%	85.09%	44.19%	58.16%	82.73%	55.68%	65.67%
6	**80.94**%	65.20%	72.23%	**85.83**%	42.32%	56.69%	**83.38**%	53.76%	64.46%

TABLE 6.4. Average measures for a particular number of languages in a combination (from one through six) for meta-classifiers.

I experiment with a majority vote meta-classifier, which combines the predictions of the *monolingual* Naïve Bayes classifiers described in Section 6.2.2.1. For a particular number of languages (one through six), all possible combinations of languages are considered. Each combination suggests a prediction only if its component classifiers agree, otherwise the system returns an *unknown* prediction. The averages are computed across all the combinations featuring the same number of languages, regardless of language identity.

The results are shown in Table 6.4. The macro precision and recall averaged across groups formed using a given number of languages are presented in Figure 6.9. If the average monolingual classifier has a precision of 74.07%, the precision increases as more languages are considered, with a maximum precision of 83.38% obtained when the predictions of all six languages are considered (56.02% error reduction). It is interesting to note that the highest precision meta-classifier for groups of two languages includes German, while for groups with more than three languages, both Arabic and German are always present in the top performing combinations. English only appears in the highest precision combination for one, five and six languages, indicating the fact that the predictions based on Arabic and German are more robust.

Let us further analyze the behavior of each language considering only those meta-classifiers that include the given language. As seen in Figure 6.10, all languages experience a boost in performance as a result of paired language reinforcement. Arabic gains an absolute 11.0% in average precision when considering votes from all languages, as compared to the 72.30% baseline consisting of the precision of the classifier using only monolingual features; this represents an error reduction in precision of 66.71%. The other languages experience a similar boost, including English which exhibits an error reduction of 50.75% compared to the baseline. Despite the fact that

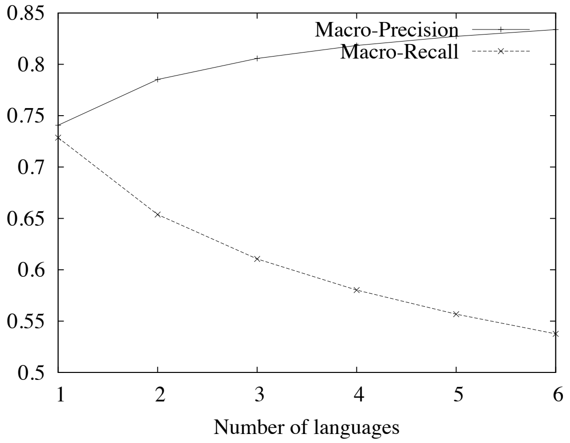


FIGURE 6.9. Average macro-precision and recall across a given number of languages.

with each language that is added to the meta-classifier, the recall decreases, even when considering votes from all six languages, the recall is still reasonably high at 53.76%.

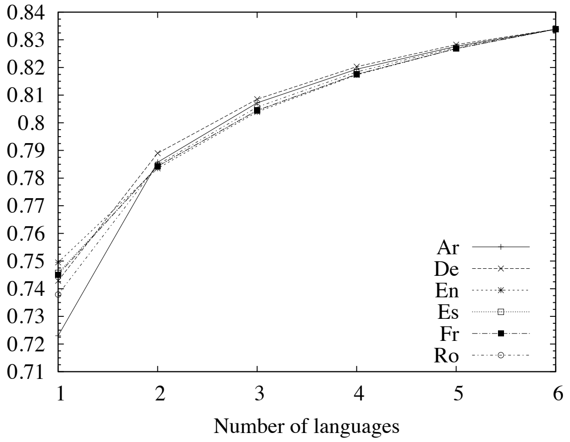


FIGURE 6.10. Average macro-precision relative to a given language.

The results presented in table 6.4 are promising, as they are comparable to the ones obtained in previous work. Compared to [80], who used a high-precision rule-based classifier on the English MPQA corpus (see Table 3.1), the multilingual method has a precision smaller by 3.32%, but a recall larger by 21.16%. Additionally, unlike [80], which requires language-specific rules, making it applicable only to English, the method described here can be used to construct a high-precision classifier in any language that can be connected to English via machine translation.

CHAPTER 7

MARKERS OF SUBJECTIVITY ACROSS LANGUAGES

So far, I have explored methods of extrapolating subjectivity research to other languages, that do not benefit of the same amount and quality of resources available for natural language processing in English. I now turn toward motivating the importance of developing such methods. First of all, as I mentioned in Chapter 1, only 27% of the Internet users speak English, and other languages have experienced a sustained growth in the amount of electronic text available online. This fact alone provides sufficient motivation for subjectivity research to be conducted in these languages as well. Yet, another reason behind this research direction is that other languages may be uniquely equipped to encode subjectivity in discourse, and in so doing may provide a lending hand with subjectivity detection in English and other languages as well. This possible trend was suggested through the experiments conducted in Chapters 4 and 6. Here, I will discuss a case in point represented by Romanian, a language that I am a native speaker of, and identify elements that aid in subjectivity research in this language, and that do not have an equivalent in English. I will first overview several facets of Romanian grammar and speech that can express subjectivity, and then I will qualitatively and quantitatively explore these aspects.

7.1. Expressing Subjectivity in Romanian

Romanian is a Romance language, being most similar to Italian, yet sharing a common ancestry with languages such as Spanish, Portuguese and French. As such, it follows a rather complex grammatical structure derived from Latin, such as grammatical cases (declination) and verb inflection (conjugation). Below, I identify several ways in which Romanian is able to express subjectivity, in addition to those available in English.

7.1.1. Verb Mood

Verbs are a part-of-speech category, along with nouns, adjectives, pronouns, articles, determiners, etc.; when embedded in a sentence, they are able to undertake various roles dictated by the way they are employed. Verbs may be finite or non-finite. Finite verbs play the role of a predicate that accompanies the subject of a sentence, and they are in agreement with the subject in person, gender, and number. This agreement, as well as additional nuances of the verb in the sentence (such as voice, mood, aspect and tense), are typically marked through inflections in the form of the verb, also known as conjugation. Non-finite verbs do not accomplish the role of predicate in a sentence, but rather behave as attribute, object or complement.

As a weakly inflected language, English allows for limited conjugation forms and relies instead on modeling voice, mood, aspect and tense through the usage of auxiliary verbs. In contrast, Romanian is a highly inflected language, featuring four verb classes that follow different conjugation rules, and where inflections in the verb change based on person, number, voice, mood, aspect and tense.

Let us consider the following sentence, where the superscript denotes the sentence number, the underline marks the predicate and the dashed line marks the subject:

She feels tired.[1]/

This example represents a complete sentence, as it contains the two primary parts, namely the subject and the predicate; objects and attributes are optional. The subject in this case is *she*, a personal pronoun in the third person, feminine, singular form. The predicate is *feels*, expressed through a verb in the indicative mood, which is in agreement with the subject in both person and number, as the infinitive form of the verb in English receives the *-s* suffix to match the third person singular form of the subject. *Tired* is a predicate adjective that provides additional information about the subject.

Personal moods are used to represent something that is actually the case (realis mood - see Table 7.1), not the case (irrealis mood - see Table 7.2), orders and prohibitions (imperative mood) and questions (interrogative mood). In English and other romance languages, the indicative mood is used to express an action that is certain or real. The irrealis mood, which groups together modes such as subjunctive and conditional, marks actions that are to be realized, or could be realized if some other hypothetical event occurred. The imperative and interrogative mood are self explanatory.

From all the personal moods listed above, the irrealis mood is unique in its ability of expressing subjective content. Subjunctive in particular marks ideas that are subjective or uncertain, such as emotions, doubts, opinions, judgments, etc., and it provides a unique marker for subjectivity. It typically appears in subordinate sentences. In English, the form of a verb in subjunctive does not carry any particular markers that allows for an easy recognition of this mood:

I suggest ^{1}/ that Jenny exercise several times a week.^{2}/

In this example, *exercise* is a verb in the subjunctive mood. It is not the indicative form, since Jenny is not actually exercising, but rather encodes a hypothetical wish enunciated by the speaker in regards to Jenny. The indicative form would have required that the proper agreement between the subject *she* and the verb be marked through the suffix *-s*. While English does not entail observable morphological changes in the form of the verb, in Romanian verbs in subjunctive are marked through the particle *să* that precedes the verb. This particle occurs uniquely in front of a verb in subjunctive mood. The example above becomes in Romanian:

Sugerez^{1}/ ca Jenny să facă mişcare de câteva ori pe săptămână. ^{2}/

where the dotted line marks the particle *să*. This particle allows for an easy detection of subjective content, that can be easily leveraged automatically.

Other moods that fall under the irrealis umbrella are conditional and optative, which mark a possible intended or desired action, whose accomplishment is contingent upon another action taking place (in the case of conditional). In non-technical discourse, conditional also has the ability to mark subjective content, as it is able to encode wishes, desires, speculations, beliefs, uncertainties. Let us consider the following conditional phrase:

I would eat ^{1}/if I were hungry. ^{2}/

In English, conditional is expressed only in the principal sentence, by appending the auxiliary “would” to the verb. In Romanian, in a conditional structure, both the main and the dependent clauses require the usage of the verb in its conditional form. This form is derived by juxtaposing the auxiliary of the verb “to be” conjugated for conditional and the main verb in its infinitive or past participle form. This facet may allow an automatic system to have an easier task in identifying subjective sentences in Romanian.

7.1.2. Verb Person and Number

Due to the highly inflected nature of verb conjugation in Romanian, the person and number information can easily be identified from the suffix and prefix of the predicate or the auxiliary verb. For example, see the markings in italics in the rightmost column in Tables 7.1 and 7.2. These allow for an accurate identification of the subject of the clause, ensuring that verbs about self (that may be important in subjectivity detection) are easily recognizable.

7.1.3. Formal versus Informal Register

A linguistic register represents the usage of a language that is dictated by a particular setting and scope. For example, an informal register is used in familiar settings in day to day conversations. It abounds in common words, repetitions, cumulative words such as *things, stuff*, vulgar words, pejorative expressions, verbs in imperative mood, etc.. By contrast, the written register requires respecting formal norms such as using proper grammar (with correct agreement, conjugation, declination), rich sentence structure and syntax, expanded vocabulary, usage of specific terms instead of generic words, using discourse rhetoric markers, etc.

While English is more rigorous in following this register spectrum, due to its weakly inflected nature, verbal expressions, shorter sentences, high overlap between spoken and written language vocabulary, the overlap between the formal and informal registers is significant. In Romanian, this is not the case. Written sentences are many more times longer compared to those encountered in spoken discourse; just as in English, a complex phrase is constructed from several main clauses connected with coordinating conjunctions or comma, and enriched with additional information provided by subordinate clauses preceded by subordinating conjunctions. However, word inflections allow for robust meaning expression even in extremely long phrases, trailing more than half a page (unlike in English). Furthermore, words and constructs that appear in the informal

Realis Mood **Indicative**			
Present Tense	*Present Continuous*		*Ro: Indicativ prezent*
I work	I *am* working		eu munc*esc*
You work	You are working		tu munce*şti*
He/she work*s*	He/she *is* working		el/ea munce*şte*
We work	We are working		noi munc*im*
You work	You are working		voi munc*iţi*
They work	They are working		ei/ele munc*esc*
Present Perfect	*Present Perfect Continuous*	*Past Tense*	*Ro: Perfect compus*
I have worked	I have been working	I worked	eu *am* muncit
You have worked	You have been working	You worked	tu *ai* muncit
He/she *has* worked	He/she *has* been working	He/she worked	el/ea *a* muncit
We have worked	We have been working	We worked	noi *am* muncit
You have worked	You have been working	You worked	voi *aţi* muncit
They have worked	They have been working	They worked	ei/ele *au* muncit
Past Continuous			*Ro: Imperfect*
I *was* working			eu munc*eam*
You were working			tu munc*eai*
He/she *was* working			el/ea munc*ea*
We were working			noi munc*eam*
You were working			voi munc*eaţi*
They were working			ei/ele munc*eau*
Past Perfect	*Past Perfect Continuous*		*Ro: Mai mult ca perfect*
I had worked	I had been working		eu munc*isem*
You had worked	You had been working		tu munci*seşi*
He/she had worked	He/she had been working		el/ea munc*ise*
We had worked	We had been working		noi munc*iserăm*
You had worked	You had been working		voi munc*iserăţi*
They had worked	They had been working		ei/ele munc*iseră*
Future Tense	*Future Continuous*		*Ro: Viitor*
I will work	I will be working		eu *voi* munci
You will work	You will be working		tu *vei* munci
He/she will work	He/she will be working		el/ea *va* munci
We will work	We will be working		noi *vom* munci
You will work	You will be working		voi *veţi* munci
They will work	They will be working		ei/ele *vor* munci
Future Perfect	*Future Perfect Continuous*		*Ro: Viitor anterior*
I will have worked	I will have been working		eu *voi* fi muncit
You will have worked	You will have been working		tu *vei* fi muncit
He/she will have worked	He/she will have been working		el/ea *va* fi muncit
We will have worked	We will have been working		noi *vom* fi muncit
You will have worked	You will have been working		voi *veţi* fi muncit
They will have worked	They will have been working		ei/ele *vor* fi muncit

TABLE 7.1. Conjugation of verb "to work" at indicative in English and Romanian. The right column contains a relative mapping to a Romanian tense (Note: English and Romanian do not have one to one tense mappings.). The *italic* markings in the conjugation offer information regarding the person and number.

register draw strong attention when they appear in writing, as they imply a strong personal bias on behalf of the author; this facet is explored below[1].

[1] The newspaper quotes were taken from Krieb Stoian, S., *Mijloace lingvistice de exprimare a aproximării în presa scrisă actuală (Linguistic means for expressing approximation in written contemporary press)*. `http://ebooks.unibuc.ro/filologie/dindelegan/19.pdf`. The interpretation is my own.

Irrealis Moods	
Subjunctive	
Translation	*Conjunctiv prezent*
I work	eu să munc*esc*
You work	tu să munc*eşti*
He/she work	el/ea să munc*ească*
You work	noi să munc*im*
They work	voi să munc*iţi*
I had work	ei/ele să munc*ească*
Translation	*Conjunctiv perfect*
I had worked	eu să fi muncit
You had worked	tu să fi muncit
He/she had worked	el/ea să fi muncit
We had worked	noi să fi muncit
You had worked	voi să fi muncit
They had worked	ei/ele să fi muncit
Conditional	
Translation	*Condiţional prezent*
I would work	eu *aş* munci
You would work	tu *ai* munci
He/she would work	el/ea *ar* munci
We would work	noi *am* munci
You would work	voi *aţi* munci
They would work	ei/ele *ar* munci
Translation	*Condiţional perfect*
I would have worked	eu *aş* fi muncit
You would have worked	tu *ai* fi muncit
He/she would have worked	el/ea *ar* fi muncit
We would have worked	noi *am* fi muncit
You would have worked	voi *aţi* fi muncit
They would have worked	ei/ele *ar* fi muncit

TABLE 7.2. Conjugation of verb "to work" in irrealis moods in English and Romanian. The left column contains the English transation of the Romanian tense (Note: English and Romanian do not have one to one tense mappings.). The *italic* markings in the conjugation offer information regarding the person and number.

Let us consider the following excerpt from Adevărul [2]:

Ro: Nu au lipsit nici pensionarii, veniţi în special pe la orele 14.30-15 [...].

En: Also not missing were the retirees, especially arrived around 14,30-15 hours [...][3].

In this case subjectivity is expressed in multiple ways. First the subject verb inversion draws attention to the fact that the retirees were not missing, implying that they had to be there, and thus signaling irony and contempt held by the author towards them. The insertion of the adverb *nici* (literal translation in English is *neither*, translated in context as *also*), expresses a negation that is

[2] Romanian newspaper. Adevărul, 3864/2002, p.4.

[3] Though this translation seems contrived, it seeks to maintain the subjective markers appearing in the Romanian fragment.

stronger than *nu* (En: *no*) and it is further doubled by *nu* appearing, generating a highly subjective construct. The usage of all these negations prompt for the need to use a negative verb *a lipsi* (En: *to be missing*), instead of a positive one *a veni* (En: *to come, to attend*) that would have been able to plainly state that the retirees were present. Such a construct is reserved for the informal register. The second part of the sentence *veniti in special pe la orele* (En: *arrived especially around*) marks again the subjective presence of the author, as he underlines that the retirees came especially around a particular time frame prompting the reader to wonder on the untold reason for the choice of the hour.

Another excerpt from Jurnalul Naţional newspaper [4]:

> **Ro**: Supărat că jucătorii aduşi la sugestia lui Marcel Popescu pe care a plătit o căruţă de bani [...] nu sunt băgaţi în teren, Mititelu l-a încondeiat pe Cârţu.
> **En**: Upset because the players brought in at Marcel Popescu's suggestion, for whom he paid a truckload of money, are not put into the field, Mititelu disparaged Cârţu.

This example contains several contaminations from the informal register. First of all, the author's usage of the expression *o căruţă de bani* (En: *a truckload of money*) signals subjectivity, primarily because this expression is used exclusively in verbal communication, and also because it denotes a subjective judgment with regards to the large quantity of money the player was paid. Moving to the next informal register marker, the verb *a băga* (En: *to put, insert*) entails an informal use that does not occur outside of verbal expressions (such as *a băga în seama* (En: *to notice*), or *a băga în sperieţi* (En: *to scare*)) in written text.

These examples are not an exception in Romanian journalism, but rather the norm. [86] remarks that this particularity (i.e. the infiltration of elements par excellence informal into the formal register) is a unique recent development in the Romanian language. She motivates this trend on Romania's history. Until 1989, when the Romanian Revolution took place, the written and the journalistic register was extremely formal and bland due to the censorship imposed by the communist government; the written text was following an administrative / judicial report form. Writers

[4] Jurnalul Naţional, 2880/2002, p.10

considered that if they cannot freely talk about particular subjects, they should avoid them completely; the government maintained the same line. As Zafiu mentions, “censorship itself prefers omission, because in very clear cases of conflict with reality, any assertion become subversive: no matter how neutral the assertion, it is perceived as an allusion, and the eulogy appears as irony.” All this changed after 1989, when the media and all Romanians started speaking freely, and after decades of censorship, the need to express and share personal thoughts in the most familiar tone is at the highest.

These contaminant words and constructs represent subjectivity markers that are unique to the Romanian language. They could be automatically harvested by creating pre 1989 and post 1989 corpora, and thus improving subjectivity classification in Romanian.

7.1.4. Expressions of Politeness

Furthermore, Romanian is a language that uses politeness pronouns in discourse (see Table 7.3). These indicate an attitude of respect (private state), but may also signal distance, i.e. not being familiar with the person the discussion is with or about. These pronouns do not have direct translations into English.

Person	Case	Singular		Plural	
		Romanian	English	Romanian	English
2nd	N, A	dumneata, dumneavoastră	you	dumneavoastră	you
	D, G	dumitale, dumneavoastră	yours, to you	dumneavoastră	yours
3rd	N, A, D, G	dumnealui (masc.) / dumneaei (fem.) / dumneasa	him, her, his, hers	dumnealor	they, their

TABLE 7.3. Politeness pronouns in Romanian; cases: nominative (N), accusative (A), dative (D), genitive (G).

7.2. Qualitative and Quantitative Analysis

In order to assess how the above elements may participate in automatically classifying subjectivity in Romanian, I start out by creating a sentence level subjectivity dataset in this language. I use the manual translation into Romanian of the SemCor corpus, also used in Chapter 6. This corpus is annotated for subjectivity at the sentence level by a native speaker of Romanian who has participated in previous subjectivity annotation studies, thus obtaining a set of 1552 subjective and 426 objective sentences. The corpus was particularly difficult to annotate because it is written in a discourse style, namely the original author is making statements about particular topics, that range from religion, art, sports to politics. For this reason, sentences that do not have any obvious markers of subjectivity, such as:

> Since man can only live by dying, so only through this death Christ could bring many to life.
>
> The payment for sin is death.
>
> Man did not comply with these conditions, and thus became mortal; although he did not cease to be human because of what he did.

end up being subjective anyway, as they contain the opinion of the writer. Due to the high bias toward the subjective class, this dataset was supplemented with approximately 500 manually annotated objective sentences extracted from the Romanian Wikipedia[5] on the same topics and containing some of the same proper nouns as those in SemCor[6]. Despite the fact that Wikipedia is an encyclopedic resource, which should be written in a non-subjective style, it was surprising to note that in a given article from the Romanian Wikipedia only the very first sentences are objective, and that the large majority of the article is subjective, thus rendering this approach of culling objective sentences non-trivial. In the end, the final set contains 912 objective and 1566 subjective sentences. These sentences were part-of-speech tagged using the RACAI web service[7] developed by [42], which uses a tagset of 79 morpho-syntactic categories as well as an additional 8 tags dedicated for punctuation. The first category of tags appends to a traditional categorization into noun, verb, adverb, adjective, pronoun, preposition, etc., additional granularity, such as markers for singular or plural, case or person (in the case of verbs). However, since many of these were not complete (i.e. few grammatical cases and moods were marked), and others were not necessarily useful for the task of subjectivity detection, I primarily employed the lemmatized form of the word provided by the RACAI service and its traditional part-of-speech, with the following exceptions. First, in the case of verbs, the part-of-speech tagger marks the subjunctive particle *să* mentioned in

[5] ro.wikipedia.org

[6] Text containing similar proper nouns was sought in order to lessen the statistical impact of rare words in the corpus.

[7] http://www.racai.ro/webservices/TextProcessing.aspx

Section 7.1.1, whose annotation is kept. Second, in addition to a generic verb tag for all verbs, the predicates also maintain the person annotation; thus the verb *muncesc* (En: I work), conjugated in the first person singular, is added as two features, once it is annotated with the generic verb tag *V*, and once with the verb-person tag *V1* (verb conjugated at the first person). We maintain this information in order to verify whether a particular verb conjugation may be a marker for subjectivity as discussed in Section 7.1.2. Third, since the tagger identifies proper nouns, and because they should not have a real bearing on the subjectivity of a sentence, we use this tag to remove proper nouns from the corpus.

A vectorial space is built from the pairs of lemmatized Romanian words and their part-of-speech with a frequency higher than 2 using binary weighting[8]. Several additional features are added, three that mark whether a sentence contains a predicate that was conjugated in the first, second and third person, respectively, and one that marks the occurrence of a politeness pronoun, that starts with *dumnea-* (as seen in Table 7.3). The intuition is that sentences containing verbs conjugated in the first person would be subjective, since they are focused on the self. Also, the politeness pronoun that indicates respect should imply subjectivity.

The first evaluations focus on the proper representation of verb conjugation. Beside including the three attributes encoding whether the predicates in a sentence are conjugated at the first, second or third person, in one experiment I used the verb-person tag, while in the other one I just maintained the verb tag. Upon performing feature selection using information gain[9], the attribute encoding for predicates conjugated in the first person ranks number 4 in both variations, while the two attributes indicating predicates conjugated in the second or third person are not even selected. This supports the initial hypothesis that conjugation may help in identifying subjectivity. Subjective verbs have a higher ranking when using the verb tag instead of the verb-person tag (which creates a sparser space), and in the top 100 features we encounter the following verbs *trebui* (En: must), *putea* (En: can), *fi* (En: to be), *spune* (En: to say), *crede* (En: to believe), *părea* (En: to seem), *afirma* (En: to affirm), *stabili* (En: to establish), and *face* (En: to do). This suggests that

[8] I have also experimented with tf and $tf.idf$ weighting, but in line with previous research, binary weighting still performs best.

[9] Available in the Weka distribution with the default threshold.

when a limited amount of data is available, it is better to use the verb tag only, paired with a status attribute that would maintain the occurrence of a predicate conjugated at the first person in the sentence. For the subsequent evaluations, the verb tag will be used.

Analyzing the attributes retained upon performing feature selection (see Table 7.4 for the highest ranking features), we notice that in the same way the subjunctive particle *să* was retained in the feature selection experiments performed for sense level subjectivity (see Table 4.2 in Chapter 4), it is also highly reliable here, occupying the 12th rank. This shows that being able to identify irrealis verbal modes is very useful for the task of subjectivity detection, and it represents a unique way in which Romanian can provide information complementary to subjectivity classification in English. While the part-of-speech tagger does not identify verbs in conditional mode, the conjunction *dacă* (En: "if") which appears in subordinate conditional clauses is ranked number 27 in the selected features. As a large part of the corpus is written in a discourse style, we notice the proper discourse markers being selected as important features, marking the author's argument progression, such as conjunctions ("and", "or", "but", "if") and adverbs ("nonetheless", "of course").

The politeness pronouns are not selected as discriminating features. Upon a closer examination, we notice that half of the instances where they occur are labeled as objective, since the writer (a coach) is referring to the readers in a polite form, telling them how to properly exercise. Despite the fact that this text was originally in English, the Romanian translator chose to use the polite form of the pronoun "you" in Romanian. [86] mentions however that in the contemporary usage of the language the politeness pronoun is replaced by the personal pronoun in advertising and in addressing potential clients in order to interact in a closer, more familiar way. Thus, depending on the text, the politeness pronouns may still play a role in subjectivity detection.

In order to assess the performance of the vectorial space created, three Naïve Bayes classifiers were trained on the following dataset variations: 1. unlemmatized - containing unigrams in the form they originally took in the text (i.e. inflected), 2. lemmatized-POS - containing tuples word - POS, where the word is in its lemmatized form, and the POS is obtained as described above[10], 3. FS-lemmatized-POS - obtained as a result of applying feature selection (information

[10] This variation also contains the 3 attributes pertaining to the verb being conjugated for the first, second, and third person, as well as the attribute encoding for the occurrence of the politeness pronoun.

gain) on the second variation of the corpus. As we notice in Table 7.5, the accuracy of the unlemmatized dataset is 73.69%, lower than the lemmatized-POS dataset at 77.08% (paired t-test, t(2476) = 2.56, $p < 0.006$) and also lower than the FS-lemmatized-POS dataset at 78.16% (paired t-test, t(2476) = 3.65, $p < 0.0002$). Furthermore, the increase in accuracy of approximately 1% between the lemmatized-POS dataset and the one obtained after feature selection is also statistically significant (paired t-test, t(2476) = 2.18, $p < 0.02$). It is also interesting to note that the main performance gains are achieved for the objective classification, which attains an improvement in F-measure by 5.7% when using the lemmatized-POS dataset, and 7.6% after feature selection, in comparison to the unlemmatized data. Similar improvements are also obtained in the F-measure for the subjective class: 2.3% when using the second variation, and 3.1% when using the third variation. As the inflections of the words aid in the part-of-speech tagging, and thus in the correct lemmatization, the lemmatized-POS space is able to provide a denser and richer representation, thus allowing higher results to be obtained.

7.3. Conclusion

This chapter has identified several ways in which Romanian text is able to encode subjectivity, in addition to those traditionally used for subjectivity classification in English. I show that some of these reflect in automatic classification experiments and further observe that that many of the attributes identified as a result of feature selection correlate with human judgments of subjectivity. Ultimately, these types of features, unique to other languages, can aid in subjectivity research in any language when approaching the task from a multilingual perspective.

Rank	Attribute ID	*Attribute*	Translation
0.13077	1216	*si_C*	and
0.05295	1176	*sau_P*	or
0.0405	905	*nu_Q*	no
0.03276	2	*_verb_V1*	Note: verb conjugated at the first person
0.03206	193	*ca_R*	like
0.02804	1217	*siliciu_Y*	and (Note: POS tagger replaced the conjunction ”and” (Ro: şi) with the chemical element Silicon, abbreviated as Si.)
0.02222	348	*dar_C*	but
0.02062	915	*obama_NC*	Obama (Note: not recognized as proper noun)
0.01591	448	*dupa_S*	after
0.01575	465	*el_P*	he
0.01433	1328	*trebui_V*	must
0.01418	1164	*sa_Q*	Note: subjunctive particle marking the subjunctive mode in Romanian
0.01161	365	*decat_R*	than
0.01134	1091	*putea_V*	can
0.01121	546	*foarte_R*	very
0.01112	800	*mai_R*	Note: marker for comparative form of adjectives in Romanian
0.01044	1200	*senat_NC*	senate
0.0095	1316	*totusi_R*	nonetheless
0.00936	1159	*roman_A*	Romanian
0.00912	868	*mult_R*	much
0.00903	747	*la_S*	to
0.00875	525	*fi_V*	to be
0.0086	1346	*un_T*	a
0.00846	212	*care_P*	that
0.00835	2584	*prea_R*	too
0.00781	480	*eu_P*	I
0.00775	347	*daca_NC*	if
0.00751	177	*bine_R*	good
0.00728	20	*acesta_D*	this
0.00711	345	*dac_A*	Note: most probably improper lemmatization
0.00708	1248	*spune_V*	to say
0.00708	858	*mod_NC*	style
0.007	739	*joc_NC*	game
0.00649	843	*meu_P*	mine
0.00648	1103	*razboi_NC*	war
0.00645	1856	*desigur_R*	of course
0.00643	378	*democrat_A*	democrat
0.00643	111	*arbitru_NC*	arbiter
0.00618	1762	*crede_V*	to believe
0.00584	1293	*tata_NC*	father
0.00583	2661	*public_NC*	public
0.00582	914	*numit_A*	named
0.00582	559	*fotbal_NC*	soccer
0.00577	863	*mondial_A*	worldwide
0.00575	1256	*stat_NC*	state
0.00566	397	*despre_S*	about
0.00551	1229	*societate_pe_actiuni_Y*	stock company
0.00534	382	*deoarece_C*	because
0.00533	2493	*parea_V*	to seem
0.00525	1404	*zid_NC*	wall

TABLE 7.4. Top 50 features selected using information gain from the Romanian corpus manually annotated for subjectivity

Corpus	Accuracy	No. attributes	Precision	Recall	F-Measure	Class
Unlemmatized	73.69%	4254	0.656	0.599	0.626	obj
			0.778	0.817	0.797	subj
			0.733	0.737	0.734	Weighted Average
Lemmatized-POS	77.08%	3060	0.695	0.671	0.683	obj
			0.812	0.829	0.82	subj
			0.769	0.771	0.77	Weighted Average
FS-Lemmatized-POS	78.17%	750	0.705	0.698	0.702	subj
			0.825	0.83	0.828	obj
			0.781	0.782	0.781	Weighted Average

TABLE 7.5. Machine learning evaluations for the Romanian corpus manually annotated for subjectivity before and after feature selection.

CHAPTER 8

CONCLUSION

In this work, I explored various methods to transfer subjectivity information from a source language to a target language. I conclude the paper by summarizing the findings, and discussing what methods worked best and why. I also revisit the questions that prompted this research and briefly summarize the findings.

8.1. Conclusions and the Big Picture

It was surprising to find that automatically growing a subjectivity lexicon in a target language starting from a small number of manually translated seeds outperforms the automatic translation of a fully developed subjectivity lexicon in a source language. The amount of manual work entailed in constructing from scratch such resources is enormous, as it requires not only training annotators on identifying subjective entities in a language, but also manually tagging a large data set, and based on the occurrence of the subjective context decide whether it is reliable enough to be included in a specialized lexicon. Compared to this method, translating a small number of seeds is a trivial task. Furthermore, obtaining an explicative dictionary in the target language and a raw corpus that can be used for extracting similarity information should not pose significant problems. Upon implementing such a bootstrapping system, we can expect up to 7% improvement in the overall F-measure over the method based on direct translation. It is also interesting to observe that despite the fact that the bootstrapping method is almost completely unsupervised (except for the initial translation of the seed set), its results are competitive with those obtained by the machine learning methods.

That is not to say that leveraging a manually annotated lexicon in a source language does not have potential. Should the subjectivity lexicon entries be placed into their afferent subjective context and this context be machine translated (as mentioned in Section 6.1.3), the entries may be properly disambiguated. Of course, such method would require textual alignment between the source and the target language translations, and the extraction of the candidate translation for the subjective entry. Yet, should such method be implemented, we can expect the emergence of a more robust subjectivity lexicon mirroring the source language resource.

While translating words requires dealing with resolving their ambiguity, using a multilingual sense aligned resource (such as the existing wordnets available in a variety of languages), is a way to bypass this shortcoming and achieve a significantly less ambiguous mapping of subjective senses across languages. As shown in the manual annotation study presented in Chapter 4, we can expect up to 90% of the senses to retain their subjective labeling after transfer to another language. This approach requires however a sense-level lexicon annotated for subjectivity in the source language, which is significantly more difficult to obtain compared to the traditional word lexicon. This type of resource is bound to have a lower coverage, and can be derived harder through automatic means. Furthermore, a significant overlap between synsets across the source and target language resource is needed, so that sufficient synsets are labeled in the target language through projection. In case automatic methods of labeling the remaining target language synsets are sought, the lexical resource in the target language has to be even richer, providing a definition or examples in addition to the synset. The sense level lexicon can then be used in the target language either by flattening its senses, and assigning a subjectivity score at the word level (as [21] have done), and then employing it as a word-level lexicon, or by measuring the similarity or overlap between a given context containing the word and the definition / examples provided with each of its senses. In this latter scenario, the word in context would receive the subjectivity label of the sense it most closely resembles. Machine learning or rule-based techniques could then be used to classify for subjectivity at the sentence, paragraph, or document level.

The machine learning experiments I proposed suggest that the use of a manual or a machine translated parallel text annotated with an automatic subjectivity analysis tool in the target language provide similar results (within 2% overall F-measure variation for both Romanian and Spanish). This furthers the notion that all the features that could have been leveraged from the automatic annotations are already taken into consideration by the machine learning algorithm. Therefore, potential for improvement is mainly entailed in the experiment focused on the machine translation of a text manually annotated for subjectivity (first experiment). The results in Spanish and Romanian adduce that should a better quality machine translation be available, additional subjectivity content can be extracted, and we should be able to better leverage the source language annotations. One possible way to surpass the need for a better translation, however, is to allow each language to work in synergy with one another. The machine learning experiments conducted following the framework of the first experiment in six languages (English, Spanish, French, German, Arabic, and Romanian) showed that when subjectivity clues arrived from multiple languages at the same time, the subjective content received a stronger contour, and was able to permeate the language boundaries, thus allowing sentences to be labeled with a higher accuracy. All languages experienced improvements over the monolingual baselines as more languages participated in the decision process. The improvements ranged from 15.27% error reduction for Arabic, to 4.90% error reduction for English.

It is also interesting to note that the best results obtained using the projection methods are only a few percentages below those that have been previously reported for experiments on English with large manually annotated data [40]. This suggests that the crafting of manual resources to obtain performance figures similar to these can be much more time consuming and expensive than using such projection methods.

Ultimately, the fact that various languages express subjectivity through different means (highlighted in our case in point study of Romanian in Chapter 7) should encourage approaches that rely on a multilingual perspective. This way, these individual aspects can be leveraged jointly, and enable superior results to those achievable when considering each language individually. In all experiments conducted here at both sense and sentence level, significant gains were seen when using a multilingual perspective (up to 77% accuracy for sense level subjectivity and up to 76% accuracy for sentence level subjectivity).

When faced with a new language, what is the best method that one can use to create a subjectivity analysis tool for that language? The answer largely depends on the monolingual resources and tools that are available for that language, e.g., dictionaries, large corpora, natural language processing tools, and/or the cross-lingual connections that can be made to a major language[1] such as English, e.g., bilingual dictionaries or parallel texts. Of course, the quality/coverage of the resources also has a bearing on how the results will compare to those we have obtained here. For languages with very scarce electronic resources, options other than translating a subjectivity lexicon may not be feasible. I am encouraged however, by the work conducted by [33] applying many of the methods presented here to new languages such as Korean, Chinese and Japanese, and obtaining comparable results. This supports the Big Picture crayoned here, at least for the more common languages.

8.1.1. Best Scenario: Manually Annotated Corpora

The best scenario is when a corpus manually annotated for subjectivity exists in the target language. Unfortunately, this is rarely the case, as large manually annotated corpora exist only for a handful of languages, e.g., the MPQA corpus that is available for English [80].

Once a large annotated data set is available, a tool for automatic annotation can be easily constructed by training a machine learning system. The task can be thus regarded as a text classification problem, and learning algorithms such as Naïve Bayes, decision trees, or SVM,[2] can be used to annotate the subjectivity of new text.

8.1.2. Second Best: Corpus-based Cross-Lingual Projections

The second best option is to construct an annotated data set by doing cross-lingual projections from a major language that has such annotations readily available. This assumes that a "bridge" can be created between the target language and a major language such as English, in the form of parallel texts constructed via manual or automatic translations. By using this bridge, the corpus annotations available in the major language can be automatically transferred into the target language. This method is described in Chapter 6.

The translation can be performed in two directions. First, one can take a collection of texts in the major language and manually or automatically translate it into the target language. In

[1] I.e., a language for which many resources and tools are already available

[2] Usually available in off-the-shelf packages such as Weka [23].

this case, if the source text is already manually annotated for subjectivity or (e.g., MPQA), then the manual labels can be projected into the target language. Alternatively, the text in the major language can be automatically annotated by using subjectivity analysis tools such as OpinionFinder [82]. The other option is to start with texts in the target language and translate them into the major language. Again, the translation can be done either by hand, or by using a machine translation system.

Regardless of the direction of the translation, and regardless of the use of manually created parallel corpora or machine translated text, the result is a data set in the target language annotated for subjectivity, which can be used to train an automatic classifier as described in the previous section.

8.1.3. Third Best: Bootstrapping a Lexicon

There are several methods that rely on the availability of subjectivity lexicons to build rule-based classifiers for the annotation of new text. For instance, one of the most frequently used subjectivity annotation tools for English is OpinionFinder [82], which is based on a large subjectivity lexicon [83].

One of the most successful approaches for the construction of subjectivity lexicons is to bootstrap from a few manually selected seeds. The bootstrapping can be performed using the synonyms and definitions found in an electronic dictionary, as illustrated in Section 5.2. No advanced language processing tools are required for this method, only a dictionary in the target language. Starting with a set of seeds covering all open-class words, all the related words found in the dictionary are collected, including the synonyms, antonyms, and words found in the definitions. From this set of candidates, only those that are closely related to the seeds are kept for the next bootstrapping iteration, with the relatedness being measured with a similarity metric. Running the process for several iterations can result in large lexicons with several thousands entries.

Sense-level subjectivity lexicons can also be bootstrapped, however they are more resource intensive compared to their word-level counterparts. They require a sense-level subjectivity lexicon in the source language and a sense-aligned bi-lingual dictionary, which can be crafted from two wordnet versions exhibiting sufficient overlap (i.e. sense *a* is mirrored by its translation *a'* in the target language) and completeness (i.e. the target language resource also provides the definition and usage examples for the sense *a'*). First, seed synsets are obtained in a target language through projections using the sense-aligned lexical resource. Then, cross-lingual or multilingual bootstrapping can be employed to label additional senses in the target language. In the experiments conducted here, the multilingual learning outperformed the cross-lingual learning, suggesting that a learner is able to make a more robust decision on how to use a multilingual space, achieving an average accuracy of 73% over 20 iterations and labeling 2000 additional synsets.

8.1.4. Fourth Best: Translating a Lexicon

If none of the previous methods is applicable to the target language, the last resort is to construct a lexicon by automatically translating an already existing lexicon from a major language. The only requirements for this approach are a subjectivity lexicon in a source language, and a bilingual dictionary used to automatically translate the lexicon into the target language. The method is described in Section 5.1. Although very simple and efficient (a lexicon of over 5,000 entries can be created in seconds), the accuracy of the method is rather low, mainly due to the challenges that are typical to a context-free translation process: difficulty in selecting the most appropriate translation for words that are ambiguous; small coverage for phrase translations; mismatch between the inflected forms appearing in the lexicon and the lemmatized forms from the bilingual dictionary. Even so, a lexicon constructed this way can be corrected by hand, and may therefore provide a building block for the generation of subjectivity resources in a given target language.

8.2. Contributions

In the remaining paragraphs, I provide a brief answer to the initial questions formulated in Chapter 1.

i. **Can subjectivity research be carried out in languages other than English without requiring in language resources?**

   In Chapters 4, 5 and 6, I explored several approaches to conducting subjectivity research in languages other than English that require no subjectivity resources to be available in the target language. The results are very promising. Up to 90% of the subjective senses maintain their subjective content upon translation. Using these annotated senses to train machine learners, allows at least six times as many senses to be labeled at an accuracy of approximately 73%. At the word level, bootstrapping a subjectivity lexicon in a target language achieves an F-measure of 64% for both Romanian and Spanish[3]. At the sentence level, using projections from a major language, allows monolingual classifiers in Spanish and Romanian to attain an F-measure above 64%. However, the most promising results are obtained when using subjectivity clues from a multilingual feature space, and learning from multiple languages at the same time. In this situation, the accuracy surpasses 75%.

ii. **Is there a benefit for subjectivity research in considering several languages at the same time? Would a multilingual model of subjectivity be more robust when compared to a traditional monolingual approach?**

In Chapters 4 and 6, I focused on setups that involve multilingual vectorial spaces and contrasted their subjectivity modeling ability against the more traditional monolingual spaces. The trend transpiring the various experiments conducted provides strong support for considering multiple languages at the same time. For the sense level subjectivity annotations, the accuracies have improved as a result of using English and Romanian in a cross-lingual or multilingual setup: from 67% for English and 70% for Romanian (monolingual baselines) to over 77% when using the multilingual setup. Even more telling are the sentence level machine learning experiments, where decisions are taken based on up to 6 languages, and where error reductions ranging from 4.9% for English to 15.27% for Arabic are obtained. This last example additionally shows that despite the fact that the original text was in English, and the subjectivity annotations were carried out on English, even this language benefited from the help provided by the other less electronic resource rich languages. As seen in Chapter 7, where I analyzed the potential of Romanian to encode subjective content in ways that are unique in comparison to English, some languages may offer subjectivity markers that are easier to detect in a machine learning setup. Thus, carrying out subjectivity research using multilingual models proves to be more robust.

[3] 5th bootstrapping iteration, 0.5 LSA threshold, seed variation

iii. **Is subjectivity a language independent phenomenon, that is able to permeate language boundaries?**

From the experiments conducted here, subjectivity seems to emerge as a language independent phenomenon. First, in the sense-level annotation study, the vast majority of senses maintained the same subjectivity judgment across languages. This aspect was able to be leveraged in a bootstrapping setup, allowing sense level classifiers to label senses with an average accuracy of 73%. Second, experiments showed that using translations of the same sentences into multiple languages allows a classifier to make stronger and more accurate subjectivity predictions. Ultimately a *private state*, whether a belief, opinion, sentiment, evaluation, or judgment tends to remain the same no matter in what language it is translated. This happens because *private states* are not a feature of discourse, but of human beings, who use language to communicate their most intimate thoughts. The sentence *A mother loves her child.* remains subjective irrespective of the underlying language in which the message is transmitted. Nuances may be more difficult to express (from where the expression *lost in translation*), but the overall expression of private state should not suffer significantly.

iv. **Is a multilingual (or monolingual) dictionary sufficient in porting subjectivity lexica to a target language?**

As shown in Chapter 5, translation using a multilingual dictionary is error prone, and attains approximately 56% F-measure for Romanian and 58.5% for Spanish . However, a resource obtained in such a way can be easily corrected by hand. More promising is growing a subjectivity lexicon using in language resources such as a monolingual dictionary. These can achieve an F-measure of 64%.

v. **Are machine translation systems able to transfer the subjective content from one language to another?**

Looking at the difference in subjectivity performance between using human translated parallel text versus text resulted from machine translation, we notice that the results are not significantly better especially when the training set is automatically annotated for subjectivity. The difference is approximately 2% for Romanian, and we can most probably expect similar results for languages that have a similar amount of electronic resources. The gap may increase however, particularly if the statistical translation engine lacks corpora in the target language.

vi. **Can we conduct multilingual subjectivity research at different granularities, be it the sense level, word level, or sentence level?**

As shown in Chapters 4, 5 and 6, multilingual subjectivity research can be carried out at different granularities. Particularly promising are the corpus based approaches and translating a sense-level annotated subjectivity lexicon.

vii. **Are there additional markers of subjectivity that appear in some languages, but not in others? What would be some unique markers of subjectivity in Romanian?**

As shown in the case in point study of Romanian in Chapter 7, a language may offer additional ways of expressing subjectivity in comparison to English. In Romanian, these range from irrealis verbal moods, to different conjugation endings pertaining to person and number, to formal versus informal registers and the way politeness is marked. These judgments correlate well with automatic feature selection, indicating that learners accurately seek to replicate human decision process.

BIBLIOGRAPHY

[1] Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, pages 190–199, Suntec, Singapore, 2009. Association for Computational Linguistics.

[2] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, pages 579–586, Vancouver, BC, Canada, 2005.

[3] Alina Andreevskaia and Sabine Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 209–216, Trento, Italy, 2006.

[4] Krisztian Balog, Gilad Mishne, and Maarten De Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, 2006.

[5] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Learning Resources Evaluation Conference (LREC 2008)*, Marrakech, Morocco, 2008.

[6] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual subjectivity: Are more languages better? In *Proceedings of the International Conference on Computational Linguistics (COLING 2010)*, pages 28–36, Beijing, China, 2010.

[7] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Sense-level Subjectivity in a Multilingual Setting. In *Proceedings of the IJCNLP 2011 Workshop on Sentiment Analysis where AI meets Psychology (IJCNLP-SAAIP 2011)*, Chiang Mai, Thailand, 2011.

[8] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Porting multilingual subjectivity resources across languages. *IEEE Transactions on Affective Computing*, page forthcoming, 2013.

[9] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Sense-level subjectivity in a multilingual setting. *Computer Speech and Language*, page forthcoming, 2013.

[10] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 127–135, Honolulu, HI, USA, 2008.

[11] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2008)*, Seattle, WA, USA, 2008.

[12] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT 1998)*, pages 92–100, Madison, WI, USA, 1998. Morgan Kaufmann.

[13] Giuseppe Carenini, Raymond T Ng, and Xiaodong Zhou. Summarizing emails with conversational cohesion and subjectivity. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*, pages 353–361, Columbus, OH, USA, 2008.

[14] Paula Carvalho, Luís Sarmento, Jorge Teixeira, and M J Silva. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics Human Language Technologies (ACL-HLT 2011)*, pages 564–568, Portland, OR, USA, 2011. Association for Computational Linguistics.

[15] Wei Chen. Dimensions of subjectivity in natural language. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (ACL-HLT 2008)*, pages 13–16, Columbus, OH, USA, 2008. Association for Computational Linguistics.

[16] Hoa Trang Dang, Diane Kelly, and Jimmy Lin. Overview of the TREC 2007 question answering track. In *Text REtrieval Conference (TREC 2007)*, 2007.

[17] Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 1988)*, pages 281–285, Washington, DC, USA, 1988.

[18] Andrea Esuli and Fabrizio Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, volume 2, pages 193–200, 2006.

[19] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 417–422, Genova, Italy, 2006. Citeseer.

[20] Andrea Esuli and Fabrizio Sebastiani. PageRanking WordNet Synsets: An Application to Opinion Mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational*, pages 424–431. Association for Computational Linguistics, 2007.

[21] Andrea Esuli and Fabrizio Sebastiani. Enhancing Opinion Extraction by Automatically Annotated Lexical Resources. In Vetulani and Zygmunt, editors, *Human Language Technology. Challenges for Computer Science and Linguistics. Lecture Notes in Computer Science*, volume 6562, pages 500–511. Springer Berlin / Heidelberg, 6562/2011 edition, 2011.

[22] R. Ghani, R. Jones, and D. Mladenic. Using the Web to create minority language corpora. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2001.

[23] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[24] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics (EACL 1997)*, pages 174–181, Madrid, Spain, 1997.

[25] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD 2004)*, pages 168–177, Seattle, WA, USA, 2004.

[26] Yi Hu, Jianyong Duan, Xiaoming Chen, Bingzhen Pei, and Ruzhan Lu. A new method for sentiment classification in text retrieval. In *Proceedings of the Second international joint conference on Natural Language Processing (IJCNLP 2005)*, Lecture Notes in Computer Science, pages 1–9, Jeju Island, South Korea, 2005.

[27] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML 1998)*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142, Chemnitz, Germany, 1998. Springer-Verlag.

[28] Nobuhiro Kaji and Masaru Kitsuregawa. Automatic construction of polarity-tagged corpus from HTML documents. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 452–459, Sydney, Australia, 2006.

[29] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1075–1083, Prague, Czech Republic, 2007. Computational Linguistics.

[30] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 355–363, Sydney, Australia, 2006.

[31] Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 494–500, Geneva, Switzerland, 2004.

[32] Noriko Kando, Teruko Mitamura, and Tetsuya Sakai. Introduction to the NTCIR-6 Special Issue. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(2):4:1–4:3, 2008.

[33] Jungi Kim, Jin-Ji Li, and Jong-Hyeok Lee. Evaluating multilanguage-comparability of subjectivity analysis systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 595–603, Uppsala, Sweden, 2010.

[34] Soo-Min Kim and Eduard Hovy. Identifying and analyzing judgment opinions. In *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*, number 2003, pages 200–207, New York, NY, USA, 2006.

[35] Nozomi Kobayashi, Kentaro Inui, Kenji Tateishi, and Toshikazu Fukushima. Collecting evaluative expressions for opinion extraction. In *Proceedings of the First international joint conference on Natural Language Processing (IJCNLP 2004)*, pages 596–605, Sanya, China, 2004.

[36] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[37] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[38] Baoli Li, Yandong Liu, Ashwin Ram, Ernest V. Garcia, and Eugene Agichtein. Exploring question subjectivity prediction in community QA. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*, pages 735–736, Singapore, Singapore, 2008. ACM Press.

[39] Jun Li and Maosong Sun. Experimental study on sentiment classification of Chinese review using machine learning techniques. In *2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 393–400, Beijing, China, August 2007. IEEE.

[40] Yaoyong Li, Kalina Bontcheva, and Hamish Cunningha. Experiments of opinion analysis on two corpora MPQA and NTCIR-6. In *Proceedings of the Sixth NTCIR Workshop Meeting*, pages 323–329, Tokyo, Japan, 2007.

[41] Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia : A system for large-scale news analysis. In *String Processing and Information Retrieval*, volume 3772 of *Lecture Notes in Computer Science*, pages 161–166. Springer Berlin Heidelberg, Berlin / Heidelberg, 2005.

[42] Oliver Mason and Dan Tufi. Probabilistic tagging in a multi-lingual environment: Making an English tagger understand Romanian. In *Proceedings of the 3rd International TELRI Seminar: Translation Equivalence - Theory and Practice*, pages 165–168, October 1997.

[43] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 976–983, Prague, Czech Republic, 2007.

[44] G. Miller, C. Leacock, T. Randee, and R. Bunker. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, Plainsboro, New Jersey, 1993.

[45] George A. Miller. WordNet: a Lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41, 1995.

[46] Smruthi Mukund and Rohini K. Srihari. A vector space model for subjectivity classification in Urdu aided by co-training. In *Proceeding of Coling 2010: Poster Volume (COLING 2010)*, number August, pages 860–868, Beijing, China, 2010.

[47] F. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, October 2000.

[48] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 271–278, Barcelona, Spain, 2004.

[49] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global WordNet (GWN 2002)*, Mysore, IN, USA, 2002.

[50] Guillaume Pitel and Gregory Grefenstette. Semi-automatic building method for a multidimensional affect dictionary for a new language. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

[51] Ana-Maria Popescu, Oren Etzioni, and Bao Nguyen. Extracting Product Features and Opinions from Reviews. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 339–346, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.

[52] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 105–112, Sapporo, Japan, 2003.

[53] Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 1106–1111, Pittsburgh, PA, USA, 2005. AAAI Press.

[54] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. In *Readings in Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, CA, 1997.

[55] Swapna Somasundaran, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and news. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, Boulder, CO, USA, 2007.

[56] Philip J. Stone, Marshall S. Smith, Daniel M. Ogilivie, and Dexter C. Dumphy. *The General Inquirer: A computer approach to content analysis*. /. The MIT Press, 1st edition, 1966.

[57] Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In Yan Qu, James Shanahan, and Janyce Wiebe, editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Menlo Park, CA, USA, 2004. AAAI Press.

[58] Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text ANN. In *Proceedings of the 4th International Workshop on the Semantic Evaluations*, 2007.

[59] Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources (LREC 2004)*, pages 1083–1086, Lisbon, Portugal, 2004.

[60] Fangzhong Su and Katja Markert. Eliciting subjectivity and polarity judgements on word senses. In *Proceedings of the 22nd International Conference on Computational Linguistics Workshop on Human Judgements in Computational Linguistics*, number August, pages 42–50, Manchester, UK, 2008. Association for Computational Linguistics.

[61] Fangzhong Su and Katja Markert. From Words to Senses: A Case Study of Subjectivity Recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 825–832, Manchester, UK, 2008.

[62] Fangzhong Su and Katja Markert. Subjectivity recognition on word senses via semi-supervised mincuts. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, number 2006, pages 1–9, Boulder, CO, USA, 2009.

[63] Fangzhong Su and Katja Markert. Word sense subjectivity for cross-lingual lexical substitution. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 357–360, Los Angeles, CA, USA, 2010.

[64] Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. Application of Semi-supervised Learning to Evaluative Expression Classification. In *Proceedings of the 7th International Conference on Intelligent Text (CICLing 2006)*, pages 502–513, Mexico City, Mexico, 2006.

[65] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting Semantic Orientations of Words using Spin Model. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 133–140, Ann Arbor, MI, USA, 2005.

[66] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 201–208, Trento, Italy, 2006.

[67] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia, 2006. Association for Computational Linguistics.

[68] Benjamin K Y Tsou, Raymond W M Yuen, Oi Yee Kwong, Tom B Y Lai, and Wei Lung Wong. Polarity Classification of Celebrity Coverage in the Chinese Press. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[69] Dan Tufi, Verginica Mititelu Barbu, Luigi Bozianu, and Ctlin Mihil. Romanian Wordnet: current state, new developments and applications. In *Proceedings of the 3rd Conference of*

the Global WordNet Association (GWC'06), pages 337–344, Seogwipo, Republic of Korea, 2006.

[70] P. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany, 2001.

[71] Peter D Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424, 2002.

[72] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[73] Xiaojun Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu, HI, USA, 2008.

[74] Xiaojun 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 (ACL-IJCNLP-2009)*, volume 1, pages 235–243, Suntec, Singapore, 2009.

[75] Janyce Wiebe, Rebecca Bruce, and Thomas O'Hara. Development and use of a gold standard data set for subjectivity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999)*, pages 246–253, College Park, MD, USA, 1999.

[76] Janyce Wiebe and Rada Mihalcea. Word sense and subjectivity. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006)*, pages 1065–1072, Sydney, Australia, 2006.

[77] Janyce Wiebe and Ellen 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 2005)*, pages 486–497, Mexico City, Mexico, 2005.

[78] Janyce Wiebe and Ellen Riloff. Finding mutual benefit between subjectivity analysis and information extraction. *IEEE Transactions on Affective Computing*, 2(4):175–191, October 2011.

[79] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, and Matthew Bell. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

[80] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[81] Edwin B. Wilson. Probable inference , the law of succession , and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[82] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, Vancouver, BC, Canada, 2005.

[83] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 347–354, Vancouver, BC, Canada, 2005.

[84] Y. Yang and X. Liu. A reexamination of text categorization methods. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[85] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 129–136, Sapporo, Japan, 2003.

[86] Rodica Zafiu. *Diversitate stilistica in romana actuala (Stylistic diversity in contemporary Romanian)*. Editura Universitatii, Bucharest, 2001.

[87] Taras Zagibalov and John A. Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on*

Computational Linguistics (COLING-2008), volume 1, pages 1073–1080, Manchester, UK, 2008.

[88] Tom Chao Zhou, Xiance Si, Edward Y Chang, Irwin King, and Michael R Lyu. A data-driven approach to question subjectivity identification in community question answering. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pages 164–170, Toronto, Ontario, Canada, 2012.