Sense-level Subjectivity in a Multilingual Setting

Carmen Banea[a], Rada Mihalcea[a], Janyce Wiebe[b]

[a]*Department of Computer Science and Engineering*
University of North Texas, Denton, TX
[b]*Department of Computer Science*
University of Pittsburgh, Pittsburgh, PA

Abstract

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. In this paper, we seek to explore whether senses aligned across languages exhibit this trait consistently, and if this is the case, we investigate how this property can be leveraged in an automatic fashion. We first conduct a manual annotation study to gauge whether the subjectivity trait of a sense can be robustly transferred across language boundaries. An automatic framework is then introduced that is able to predict subjectivity labeling for unseen senses using either cross-lingual or multilingual training enhanced with bootstrapping. We show that the multilingual model consistently outperforms the cross-lingual one, with an accuracy of over 73% across all iterations.

Keywords: Sentiment and text classification, multilingual subjectivity analysis, sense level subjectivity

Email addresses: `carmen.banea@gmail.com` (Carmen Banea), `rada@cs.unt.edu` (Rada Mihalcea), `wiebe@cs.pitt.edu` (Janyce Wiebe)

1. Introduction

Sentiment and subjectivity analysis seeks to automatically identify opinions, beliefs, speculations, emotions, sentiments and other *private states* in natural text (Wiebe et al., 2005). Quirk et al. (1985) define a private state as a state that does not lend itself to an objective external validation, or in other words "a person may *be observed to assert* that God exists, but not to *believe* that God exists. Belief is in this sense *private.*" (p. 1181). In the field of natural language processing, researchers have used the term subjectivity analysis to denote identifying private states in text, namely separating objective from subjective instances, while sentiment or polarity analysis further refines the subjective text into positive, negative or neutral.

Sentiment and subjectivity analysis has stemmed into a prolific area of research, mainly due to the fact that numerous text processing applications stand to gain from incorporating sentiment dimensions into their models, including automatic expressive text-to-speech synthesis (Alm et al., 1990), tracking sentiment timelines in on-line forums and news (Balog et al., 2006; Lloyd et al., 2005), and mining opinions from product reviews (Hu and Liu, 2004). 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 question answering (Yu and Hatzivassiloglou, 2003), to conversation summarization (Carenini et al., 2008), text semantic analysis (Wiebe and Mihalcea, 2006; Esuli and Sebastiani, 2006) and lexical substitution (Su and Markert, 2010).

In experiments carried out on English, Wiebe and Mihalcea (2006) have shown that the most robust subjectivity delineation occurs at sense and not at word level. Following this more fine-grained perspective, Esuli and Sebastiani (2006) and Andreevskaia and Bergler (2006) have proposed methods to embed sense-level automatic sentiment annotations (objective/neutral, negative and positive) over the English WordNet structure (Miller, 1995), using its relationships (synonymy, antonymy, meronymy, etc.). On the other hand, noticing the scarcity of hand crafted sense-level subjectivity/polarity lexica, Markert and Su (2008) have explored ways to infer them from data annotated at either the word or sentence level.

Sense-level subjectivity and cross-lingual subjectivity and sentiment analysis have received considerable attentions in recent years, yet our paper explores the area that lies at the intersection of these two topics. To our

knowledge, this area has not been formally investigated, and while the techniques may be similar to those applied in sentiment and subjectivity analysis at the sentence or the review level, our work explores the more difficult task of sense-level subjectivity, which also involves deep semantic aspects of the language. The manual annotation study we performed for this task (cross-lingual sense-level subjectivity annotations), as well as the methods we proposed (cross-lingual and multilingual learning using dictionaries in multiple languages) are novel to our knowledge.

This work seeks 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 we employ a multilingual framework that can automatically discover new subjective/objective senses starting with a limited amount of annotated data? We seek to answer the first question by conducting a manual annotation study in Section 2. For the second question, we propose two models (see Section 3), one cross-lingual and one multilingual, which are able to simultaneously use information extracted from several languages when making subjectivity sense-level predictions.

2. Sense Level Subjectivity Consistency Across Languages: Annotation Study

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

We consider a sense-level aligned multilingual resource such as WordNet. WordNet (Miller, 1995) 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.

In our experiments we use the English (Miller, 1995) and the Romanian

---

[1]http://www.globalwordnet.org/gwa/wordnet_table.htm

(Tufiş et al., 2006) 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, we use sense-level manually annotated subjectivity data from (Wiebe and Mihalcea, 2006) and (Akkaya et al., 2009), as well as a list of 48 additional words, for a total of 128 words accounting for 580 English senses (with an average polysemy of 4.6). Their equivalent into Romanian is also 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 Akkaya et al. (2009) dataset) to 90% (for the Wiebe and Mihalcea (2006) dataset). When excluding senses that had both subjective and objective uses 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, we describe below in more detail the various scenarios we encountered.

In the remainder of this article, English is abbreviated as *En* and Romanian as *Ro*.

2.1. 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 1, is marked in the 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

---

[2]http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
[3]http://www.racai.ro/wnbrowser/Help.aspx

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 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.

your conclusion."[4] 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 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.

2.2. 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 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. However, in our 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.

3. Multilingual Subjectivity Sense Learning

In our previous work exploring the ability of multilingual models to better capture subjectivity at the sentence level (Banea et al., 2010), which was conducted on six languages, namely English, Arabic, German, Romanian, Spanish and French, we noticed that simultaneously considering features originating from multiple languages results in error rate reductions ranging from 5% for English to 15% for Arabic, as compared to the monolingual model baselines. The experiments also showed that the maximum improvement is achieved when the multilingual model is built over the expanded feature space comprising the vocabulary of all six

---

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

languages. This observation became the catalyst for the work presented here, 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, in this section we explore ways to use a multilingual learning mechanism to automatically predict the subjectivity of a word sense. We experiment 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, we seek to understand whether, under this scenario, a classifier is able to make a better decision by having access to the entire feature set.

We start by considering the intersection of the Romanian and English WordNets, so that we can have equivalent senses (including their definitions and synsets) in both languages. We were thus able to obtain 19,124 unique synsets. We then generate vectorial representations for two monolingual models (one in English and one in Romanian), and one multilingual model (comprising both Romanian and English features). These are composed of uni-grams extracted from a synset and its gloss, appended with a binary weight. The synset is stripped of any sense identifying features[5] 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 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>
Multilingual vector: <a$_{en}$ 1, summary 1, of 1, the 1, subject 1, or 1, plot 1,

---

[5]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.

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>

Traditionally, the subjectivity content of an entity, be it word, sentence, or document, is regarded as a binary decision (either subjective or objective). In this paper, we mimic its occurrence in natural language, and we represent it on a continuum, where 0 is at one end of the spectrum and represents full objectivity, while 1 is at the other end, and denotes full subjectivity. We establish a zone of 0.4 from the left and right of our spectrum, and we consider the synsets whose scores fall in these ranges as objective (if below 0.4) or subjective (if above 0.6). This allows us to have a buffer zone of 0.2 (above 0.4 and below 0.6), which contains 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 scores to nominal buckets), in order to run the experiments we use a linear regression algorithm,[6] which extrapolates from the data and infers a subjectivity score for every synset.

3.1. Cross-lingual Learning

The first method focuses on cross-lingual learning (CL). Based on the co-training algorithm proposed by Wan (2008), we consider the manually annotated training data in each of the languages individually, and we learn two monolingual regression algorithms (see Figure 1, (1)[7]). For every sample in the unlabeled data (2), we allow the machine learners to individually predict a score (3), and at every iteration maintain two sets with the top n most confident objective and subjective examples, respectively. These sets are ordered based on the average of 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,

---

[6]Included with the Weka machine learning distribution (Hall et al., 2009)

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

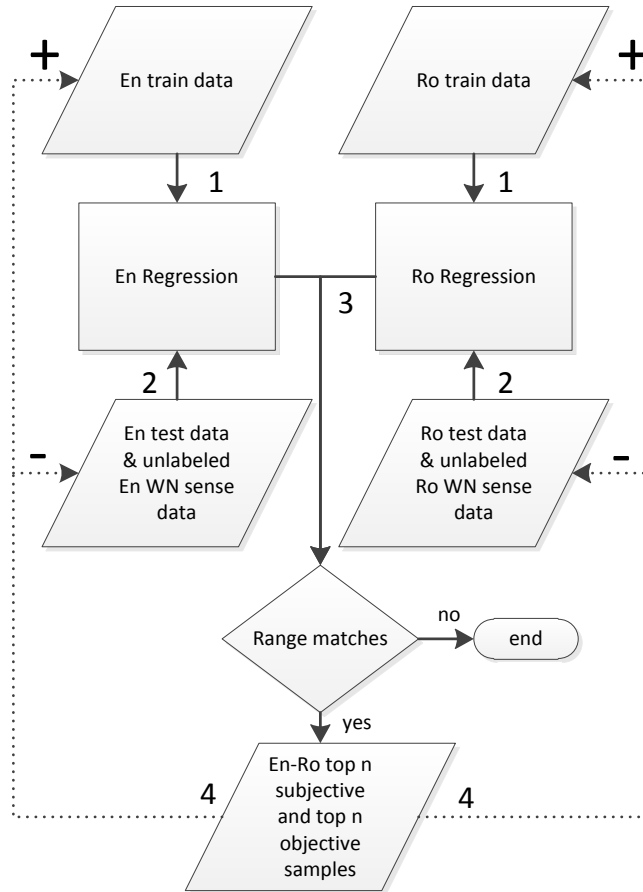and removed from their respective test set (-); otherwise the bootstrapping terminates.

Figure 1: Cross-lingual bootstrapping

Although the method differs from the original co-training mechanism proposed by Blum and Mitchell (1998), since it enforces that both predictions fall in the same range before adding the samples to the next train set, we believe 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, we ensure that only samples that have a high probability of being labeled correctly are added, therefore reducing noise propagation across iterations. At the same time, we are able to learn new information

from the features co-occurring with those that participated in the previous classification step.

3.2. Multilingual Learning

The second method employs multilingual learning (ML) (see Figure 2). We create a multilingual feature space based on the model proposed in Banea et al. (2010). **As mentioned earlier, in that work, instead of using the monolingual feature vectors to represent the sentences, we used a multilingual space combinining features drawn from up to six languages. Similarly, here,** instead of using the monolingual vectors described above, we enrich the feature space by merging together two aligned vector space representations (see the multilingual vector example above), thus allowing the system to simultaneously use both Romanian and English features in order to decide the subjectivity of a given sense. We train the multilingual learner (1) and for every sample in the testing set (2), we predict a subjectivity score (3). As we did for the cross-lingual learning setup, at every iteration we select the most confident n objective and n subjective samples (4), and add them to the training set (+), while discarding them from the test set for the next iteration (-).
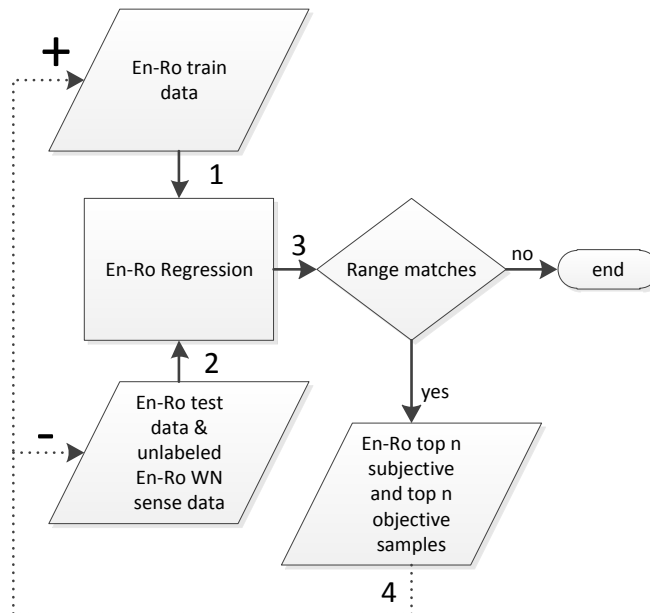

Figure 2: Multilingual bootstrapping

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 our experiments, we conducted 20 iterations for both methods and added 50 subjective and 50 objective samples at each iteration. Additional iterations would have been possible, but we decided to stop given the drop in performance of the Romanian learner embedded in the cross-lingual model.

3.3. Datasets

We use the manually annotated data described in Section 2, from which we remove 20 examples that were labeled as both objective and subjective in either English or Romanian, since they could confuse the classifiers and prevent them from making strong predictions.[8] We then split the labeled data into three subsets to enable three-fold cross validation. As these subsets are biased towards the objective class in a ratio of 2:1, we randomly discard about half of the objective samples to be included in the fold for each training set in order to obtain balanced training folds, thus allowing our experiments to not be skewed towards any of the classes. Note that we did not balance the test sets. Also, throughout every iteration the class balance is maintained as an equal number of subjective and objective samples are added to the next train set. 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, we append the remaining unlabeled WordNet senses to each test fold (see Figures 1 and 2, (2)). These running test sets are used to provide the learners with novel samples (and features) throughout the bootstrapping process. We only used 328 training examples because there is a very limited amount of subjectivity data manually annotated at the sense level in English, which moreover, needs to be mirrored in the Romanian WordNet, which has far less

---

[8]We did not remove those synsets that had conflicting labels across languages.

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, in our experiments we used all the training examples we could have for both our methods as well as the baseline, that would also allow for the existence of a small test set so that we can evaluate our results.

3.4. 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 3 is only used internally by the cross-lingual / multilingual bootstrapping methods, since its principal aim is to reduce noise propagation across iterations.

Figure 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 we notice 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 (Banea et al., 2008a), 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. Our results seem to further support the hypothesis that subjectivity analysis is an easier task in Romanian proposed by (Banea et al., 2008a). The

12

highest joint accuracy is obtained during the 4th iteration, and it represents a 3.54% improvement over the baseline.
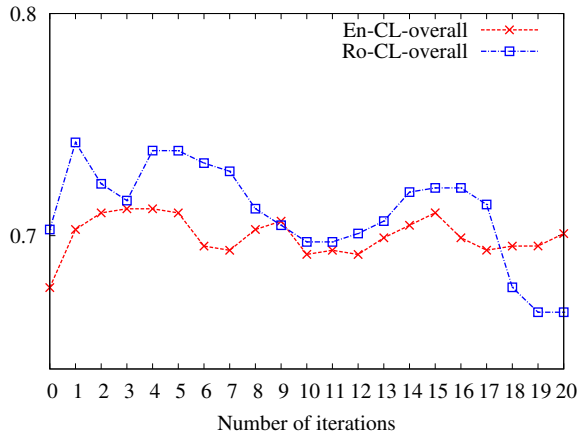

Figure 3: Macro-accuracy for cross-lingual bootstrapping

When we analyze the class behavior (see Figure 4), we notice that 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.

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), we conduct a similar experiment, this time on multilingual vectors, and allow the linear regression algorithm to provide a score. As seen in Figure 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%
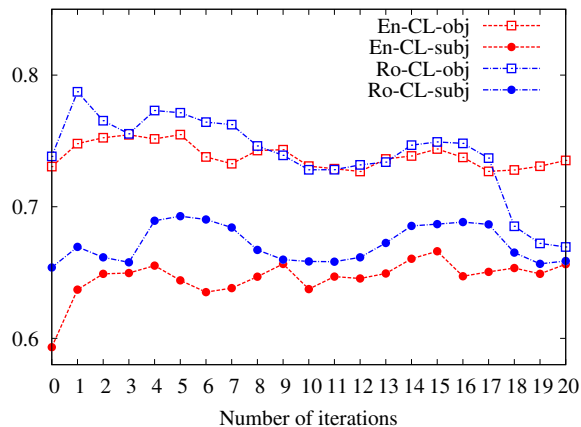
Figure 4: F-measure for the objective and subjective classes for cross-lingual bootstrapping

(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.

We also look at the class behavior under the multilingual settings (see Figure 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%. We should also note that while improvement is experienced for both the objective and subjective classes, a major gain is observed for the latter.

We are 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 Akkaya et al. (2009), where on a set of 83 English words, an accuracy of 88% was observed; and the method proposed in Su and Markert (2009), 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.
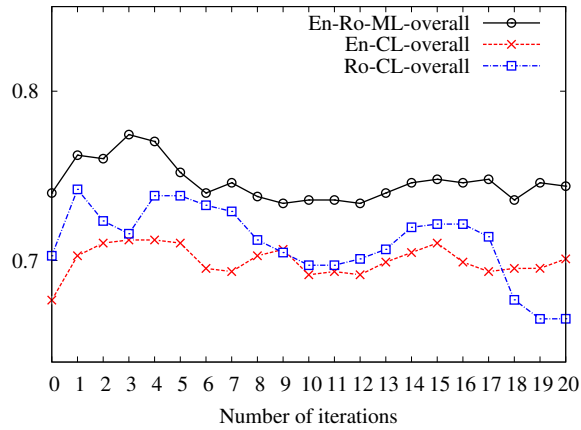
14

Figure 5: Macro-accuracy for multilingual bootstrapping (versus cross-lingual framework)
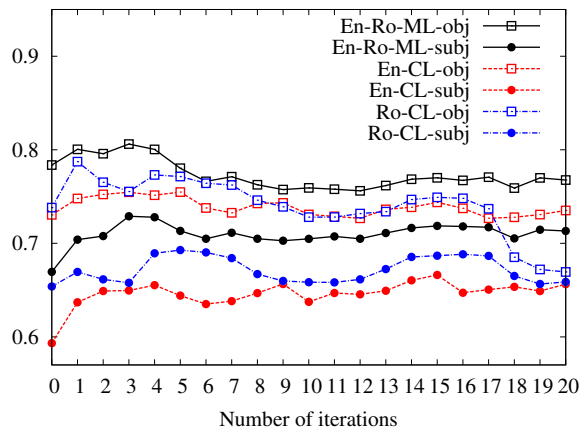

Figure 6: F-measure for the objective and subjective classes for multilingual bootstrapping (versus cross-lingual framework)

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

Table 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 multilingual features represent equivalent translations in English and Romanian.

We 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 training set generated after the tenth iteration carried out on the third fold). The subjective entries are listed in Table 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 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 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 in Figure 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 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).

---

[9]Implementation included in the Weka machine learning distribution (Hall et al., 2009)

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 used to mark ideas that are subjective or uncertain, such as emotions, doubts, opinions, judgements, etc., and it provides a unique marker for subjectivity. It typically appears in subbordinate 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 enounciated by the speaker with regards to Jenny. The indicative form would have required that the proper agreeement 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 preceeds the verb. This particle occurs uniquely in front of a verb in subjective 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*.

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 the position 29. This represents one unique example of a way in which a language provides valuable input to accurately classifying subjectivity in another language.

Our 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 our manual

annotation study in Section 2), this study strengthens 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 we were not able to conduct a study on the difference in performance when using different language pairs for sense subjectivity annotations, mainly because we did not have the required sense resources in other languages, our previous work on subjectivity at sentence level (Banea et al., 2010) seems to indicate that a more robust learning occurs when the languages are further apart. In that paper 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. Related Work

Recently, resources and tools for sentiment analysis developed for English have been used as a starting point to build resources in other languages, via cross-lingual projections or monolingual and multilingual bootstrapping. Several directions were followed, focused on leveraging annotation schemes, lexicons, corpora and automated annotation systems. English annotation schemes developed for opinionated text lays the groundwork for research carried out by Esuli et al. (2008) when annotating expressions of private state in Italian or by Maks and Vossen (2010) in Dutch. Sentiment and subjectivity lexicons such as the one included with the OpinionFinder distribution (Wiebe and Riloff, 2005), the General Inquirer (Stone et al., 1966), or the SentiWordNet (Esuli and Sebastiani, 2006) were transferred into Chinese (Ku et al., 2006; Wu, 2008), Romanian (Mihalcea et al., 2007), and Italian (Esuli and Sebastiani, 2011). English corpora manually annotated for subjectivity or sentiment such as MPQA (Wiebe et al., 2005), or the multi-domain sentiment classification corpus (Blitzer et al., 2007) were subjected to experiments in Spanish, Romanian, or Chinese upon automatic translation by Banea et al. (2008b); Wan (2008). Furthermore, tools developed for English were used to determine sentiment or subjectivity labeling for a given target language by

transferring the text to English and applying an English classifier on the resulting data. The labels were then transferred back into the target language (Bautin et al., 2008; Banea et al., 2008b). These experiments are carried out in Arabic, Chinese, French, German, Japanese, Spanish, Romanian.

Esuli and Sebastiani (2011) did consider transferring automatically inferred polarity annotations at the sense level from the English SentiWordNet to Italian using MultiWordNet (Pianta et al., 2002)[10] 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. In contrast, our work explores the ability of senses to maintain their subjectivity across languages, as well as allowing finer grained subjectivity lexical resources to be developed and enriched in other languages.

This work is related to Banea and Mihalcea (2011) who addressed the concept of subjectivity expressed at the sense level in multilingual settings, yet the results they obtained were mainly negative, and were unable to surpass the individual monolingual models using classification techniques. Our work differs as it envisions the task of subjectivity detection as a regression problem by allowing senses to be represented on a subjectivity continuum instead of a traditional binary decision. Furthermore, we show that we are able to surpass the results obtained on the original sets by implementing a bootstrapping approach.

In terms of methodology, the work closest to ours is the one proposed by Wan (2008), who constructs a polarity co-training system by using the multilingual versions obtained through the automatic translation of product reviews into Chinese and English. Unlike Wan (2008), we do not use any machine translation, and the labels employed are directly assigned by the annotators and not inferred based on stars. Furthermore, we focus on *subjectivity* classification, and not on *polarity*. Banea et al. (2008a) present a method to learn sentence level subjectivity by training classifiers on multilingual feature spaces and show that when considering features from multiple languages, the classification accuracy improves, even above that of the source

---

[10]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.

language. We expand this method to allow for bootstrapping and to be employed at the sense level, thus enabling additional samples to be classified.

5. Conclusion

We conducted a case study seeking to assess subjectivity transfer across languages following sense aligned resources. In our 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. We then proposed and applied a framework that is able to jointly exploit the subjectivity information originating from multiple languages. We demonstrated that learning from a multilingual feature space is able to capture more information and outperforms cross-lingual learning based on monolingual vectorial models, while also allowing for even better results to be obtained upon bootstrapping.

The data we developed for this work as well as the code will be made publicly available.

Acknowledgments

References

Akkaya, C., Wiebe, J., and Mihalcea, R. (2009). Subjectivity Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore.

Alm, C. O., Roth, D., and Sproat, R. (1990). Emotions from text: machine learning for text-based emotion prediction. *Intelligence.*

Andreevskaia, A. and Bergler, S. (2006). Semantic Tag Extraction from WordNet Glosses. In *Proceedings of the 5th Conference on Language Resources and Evaluation.*

Balog, K., Mishne, G., and Rijke, M. D. (2006). 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.

Banea, C. and Mihalcea, R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS'11)*, pages 25–34, Oxford, UK.

Banea, C., Mihalcea, R., and Wiebe, J. (2008a). A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Learning Resources Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Banea, C., Mihalcea, R., and Wiebe, J. (2010). Multilingual Subjectivity: Are More Languages Better? In *Proceedings of the International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008b). 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, Hawaii.

Bautin, M., Vijayarenu, L., and Skiena, S. (2008). International Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media ICWSM*, pages 19–26, Seattle, Washington.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational (ACL-2007)*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, pages 92–100. Morgan Kaufmann.

Carenini, G., Ng, R. T., and Zhou, X. (2008). Summarizing Emails with Conversational Cohesion and Subjectivity. In *Proceedings of the Associa-*

tion for Computational Linguistics: Human Language Technologies (ACL-HLT 2008), pages 353–361, Columbus, Ohio.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, IT.

Esuli, A. and Sebastiani, F. (2011). 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.

Esuli, A., Sebastiani, F., and Urciuoli, I. C. (2008). Annotating Expressions of Opinion and Emotion in the Italian Content Annotation Bank. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD-2004)*, pages 168—-177, Seattle, Washington.

Ku, L.-W., Liang, Y.-T., and Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, volume pages, pages 100–107. American Association for Artificial Intelligence.

Lloyd, L., Kechagias, D., and Skiena, S. (2005). Lydia : A System for Large-Scale News Analysis (Extended Abstract) News Analysis with Lydia. In *Lecture Notes in Computer Science*, pages 161–166. Springer, Berlin / Heidelberg.

Maks, I. and Vossen, P. (2010). Annotation scheme and gold standard for Dutch subjective adjectives. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 1327–1334, Valletta, Malta.

Markert, K. and Su, F. (2008). Eliciting Subjectivity and Polarity Judgements on Word Senses. In *Proceedings of the workshop on Human Judgements in Computational Linguistics (COLING 2008)*, pages 42—-50, Manchester, UK.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). 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.

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

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global WordNet (GWN 2002)*, Mysore, IN.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, New York.

Stone, P. J., Smith, M. S., Ogilivie, D. M., and Dumphy, D. C. (1966). *The General Inquirer: A Computer Approach to Content Analysis. /*. The MIT Press, 1st edition.

Su, F. and Markert, K. (2009). 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.

Su, F. and Markert, K. (2010). 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.

Tufis, D., Barbu, V. M., Bozianu, L., and Mihil, C. (2006). 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, Jeju Island, Republic of Korea.

Wan, X. (2008). Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, Singapore.

Wiebe, J. and Mihalcea, R. (2006). 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)*, Sydney, Australia.

Wiebe, J. and Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497, Mexico City, Mexico.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Wu, Y. (2008). *Classifying attitude by topic aspect for English and Chinese document collections*. PhD thesis, College Park, MD, USA.

Yu, H. and Hatzivassiloglou, V. (2003). 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.