# Exploring Distributional Representations and Machine Translation for Aspect-based Cross-lingual Sentiment Classification

**Jeremy Barnes  Patrik Lambert*  Toni Badia**
Universitat Pompeu Fabra, Barcelona, Spain
*Webinterpret, Barcelona, Spain
{jeremy.barnes,toni.badia}@upf.edu, patrik.l@webinterpret.com

## Abstract

*Cross-lingual sentiment classification* (CLSC) seeks to use resources from a source language in order to detect sentiment and classify text in a target language. Almost all research into CLSC has been carried out at sentence and document level, although this level of granularity is often less useful. This paper explores methods for performing aspect-based cross-lingual sentiment classification (aspect-based CLSC) for under-resourced languages. Given the limited nature of parallel data for under-resourced languages, we would like to make the most of this resource for our task. We compare zero-shot learning, bilingual word embeddings, stacked denoising autoencoder representations and machine translation techniques for aspect-based CLSC. We show that models based on distributed semantics can achieve comparable results to machine translation on aspect-based CLSC. Finally, we give an analysis of the errors found for each method.

## 1 Introduction

*Sentiment analysis* (SA) seeks to define the underlying sentiment of a text. The best results in SA require the use of a large number of resources; from tokenizers and parsers to large sentiment lexicons or hand-annotated corpora. The creation of these resources requires time, effort and a considerable monetary investment in order to ensure the quality and subsequent usefulness. Therefore, finding a way to perform sentiment analysis for under-resourced languages without having to repeat these efforts is an interesting endeavor. *Cross-lingual Sentiment Analysis* (CLSA) attempts to find methods to do just this. Most research on CLSA has been at document or sentence level. However, this does not capture the true granularity of opinionated text. Thus, in this paper we focus on CLSA at aspect-level[1]

Document- and sentence-level CLSA assume that an entire section of text expresses one sentiment towards one entity. This, however, is not always true. Aspect-based CLSA allows for multiple opinions towards multiple entities or aspects. Aspect-based CLSA can be decomposed into three subtasks: entity/aspect extraction, opinion holder extraction and sentiment classification. This last task, known as *Cross-lingual Sentiment Classification* (CLSC), has received little attention at aspect-level. Yet it would greatly benefit companies and government organizations that wish to gather information on the public opinions of their products or policies. In this paper we will only deal with improving or enabling aspect-based CLSC and leave the final goal of creating full aspect-based CLSA systems for under-resourced languages as future work.

Most research in CLSC has used Statistical Machine Translation (SMT) as a way of bridging the gap between languages, but there are drawbacks to this. First, an SMT system must be available for the language combination at hand. This requires a great deal of development and the quality of the sentiment analysis system used afterwards depends heavily on the quality of the SMT system. Secondly, study shows that even high quality SMT introduces noise into the data (Balahur and Turchi, 2014; Mohammad



[1]Here the term "aspect" refers to a feature of an entity. As an example taken from a hotel review, the sentence *"The rooms were great, but the service needs to improve"* contains two aspects (*rooms* and *service*) which pertain to the entity *hotel*. It is more useful to know that *rooms* is positive and *service* is negative than to know the overall sentiment towards *hotel*



*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan, December 11-17 2016.

et al., 2015). Finally, there are tasks in which systems which use distributed semantic representations to map between languages outperform SMT systems, e.g. cross-lingual document classification (Klementiev et al., 2012).

For this reason, a different representation of words and phrases, e.g. distributional vector representations, could prove to be a more effective approach and enable us to leverage information from resource-rich languages (English) to perform CLSA in a target language that lacks these resources (e.g. Spanish, Catalan, Basque).

This paper makes the following contributions:

- According to our knowledge, this is the most complete comparison of several types of distributed representations and machine translation for cross-lingual sentiment analysis.
- We give an analysis of the errors and possible ways to improve each system.
- We demonstrate that distributed representations can be competitive with machine translation for CLSC tasks.

## 2 Related Work

### 2.1 Monolingual Aspect-based Sentiment Analysis

*Aspect-based sentiment analysis* (ABSA) is a fine-grained approach to sentiment analysis. Many of the state-of-the-art ABSA systems in English require sophisticated NLP tools or hand-crafted sentiment lexicons. Hu and Liu (2004) propose WordNet-based methods for classifying aspect sentiment. Zhu et al. (2009) use sentiment lexicons. Moghaddam (2010) extracts an aspect and its nearest adjective and use a *k nearest neighbor* algorithm in order to estimate the rating of each aspect. Kiritchenko et al. (2014) use extracted features (part-of-speech tags, parsing features, sentiment lexicons, character-based information, n-grams) to train a *support vector machine* (SVM) for sentiment classification. These language-specific approaches do not lend themselves easily to CLSA because the target language often lacks the necessary resources.

### 2.2 Cross-lingual Sentiment Analysis

**Aspect-based CLSA**

In under-resourced languages, we lack resources and NLP tools which would allow us to create state-of-the-art systems similar to those mentioned. Therefore, the ability to leverage resources that already exist in English to perform sentiment analysis in other languages would be a great advantage. This would increase the performance of ABSA systems which are built using limited amounts of data in low-resourced languages and enable the creation of sentiment analysis systems in languages which have none at the moment.

Similarly, within CLSA, most researchers have worked at document- and sentence-level. In fact, there are only a handful of articles that deal with aspect-based CLSA. Zhou et al. (2012), Lin et al. (2014) and Klinger and Cimiano (2015) concentrate on extracting bilingual aspects but offer few ways to improve classification accuracy. Hass and Versley (2015) used machine translation and word alignment to map annotated syntactic nodes from English to German.

One of the difficulties at aspect-level is that the opinions attach to specific groupings of words, rather than a sentence or document. If we use SMT to create a new target language dataset, the opinionated units (e.g. opinion holder, opinion target, and opinion phrase) may be scattered or reordered. This would effectively reduce the usefulness of our new data because it would be difficult to project the opinion labels onto their corresponding word or phrase in the new dataset.

Lambert (2015) deals with this by using constrained SMT to translate the opinionated units within the context of the sentence. The classifiers trained on this SMT data achieve comparable results to their monolingual version. However, this is a state-of-the-art SMT system[2] which is not available in most language combinations. For these other languages, it would be useful to find alternative methods which do not require machine translation.

[2]The system achieves a *BLEU* score 45.3 in Spanish-English translation with true-case.

### CLSA via Machine Translation

Advances in machine translation have made it possible to translate data from English to a target language or vice versa and use this data to train a classifying algorithm. There are reasons to believe that for well-resourced languages machine translation has reached a level that is useful for sentiment analysis (Banea et al., 2008; Duh et al., 2011; Balahur and Turchi, 2014; Mohammad et al., 2015). Much of the work has concentrated on the best combination of translation direction, classifiers and features (Banea et al., 2008; Banea et al., 2013; Balahur and Turchi, 2014). The advantage of this approach is that it is straight-forward to use a quality SMT system to create new resources by translating annotated corpora or sentiment lexicons (Mihalcea et al., 2007).

Nonetheless, there are disadvantages to using directly translated resources. Poor translation introduces a large amount of noise which hurts the performance of the classifier (Balahur and Turchi, 2014; Mohammad et al., 2015). It is clear that under-resourced languages are particularly susceptible to poor translation. Even with high quality machine translation, there is still a cross-lingual adaptation problem; the distribution of words and their polarity do not necessarily hold in cross-lingual contexts (Guo and Xiao, 2012; Mohammad et al., 2015). Therefore, we must find ways to minimize the undesirable effects of translation in cross-lingual sentiment analysis.

### CLSA via Bilingual View

Another approach is to create a bilingual view of the data. The essence of this approach is to reduce the noise that translation introduces by presenting classifiers with complementary views. Wan (2009) creates a bilingual representation of the data through SMT and then uses co-training to take advantage of classifiers that commit complementary errors. This research seems promising, but there are some reasons to believe that the benefits of these techniques may have more to do with semi-supervised learning than cross-lingual transfer (Demirtas and Pechenizkiy, 2013).

Pan et al. (2011) use a bi-view non-negative matrix tri-factorization approach which allows for the incorporation of sentiment lexicon information. Lu et al. (2011) incorporate a joint bilingual model which makes use of unlabeled parallel or pseudo-parallel data in order to improve sentiment classification for both languages simultaneously.

### CLSA via Latent View

Zhou et al. (2016) employ stacked denoising autoencoders to create a language independent representation of their data. This representation was then used as input for a linear SVM classifier.

For cross-lingual document classification, Prettenhofer and Stein (2011) use structural correspondence learning in order to find 'pivot' features. They use these features to create a representation of each document. These latent representations that encode the relationships betweens pivots and non-pivots are then used to train a linear classifier. Klementiev et al. (2012) create bilingual distributed representations as proposed by Bengio et al. (2003). They used these word vectors to classify cross-lingual documents.

The last two techniques are not entirely comparable to the first, since they perform cross-lingual document classification and not sentiment analysis. However, approaches which use latent representations are interesting because they have the potential to avoid some of the errors which are introduced by translation. To our knowledge, many techniques for creating latent bilingual representations (Chandar et al., 2014; Gouws et al., 2015; Vulić and Moens, 2016) have not been applied to CLSA. However, these techniques could provide a straightforward way to bridge the language gap.

## 3 Methodology

### 3.1 Datasets

The data used to train the sentiment analysis models are the English and Spanish OpeNER sentiment corpora (Agerri et al., 2013). We take a subset of these corpora which deal only with hotel reviews. Each review has annotations for opinion holders, opinion targets and opinion sentiment. We refer to this triplet (opinion holder, opinion target, opinion sentiment) as an opinion unit. The sentiment can be strong positive, positive, negative, or strong negative. A neutral category is not included. As such, when

training a classifier, rather than training on the complete sentence, we use the opinion unit. Table 1 shows the statistics for these corpora.

| OpeNER Corpora | English | Spanish |
|---|---|---|
| Training Examples | 2780 | 2991 |
| Strong Pos | 23.38% | 29% |
| Pos | 46.08% | 50.34% |
| Neg | 25.61% | 17.41% |
| Strong Neg | 4.93% | 3.01% |
| Test examples | 929 | 999 |
| Strong Pos | 23.36% | 29.23% |
| Pos | 46.07% | 50.34% |
| Neg | 25.62% | 17.42% |
| Strong Neg | 4.95% | 3.00% |

Table 1: Statistics of OpeNER Corpora

The corpora used to create the word embeddings are an English and Spanish Wikipedia corpus. These were taken from Wikipedia dumps in January 2016 and preprocessed to remove html markup and lowercase all words. We then performed sentence and word tokenization. We did not remove punctuation because this is often useful information for sentiment analysis. Table 2 gives the statistics for these corpora.

| Wikipedia Corpora | English | Spanish |
|---|---|---|
| Number of sentences | 118,900,197 | 26,777,415 |
| Number of tokens | 2,055,786,401 | 506,612,108 |

Table 2: Statistics of Wikipedia Corpora

The English-Spanish part of the Europarl v7 corpus[3] (Koehn, 2005) is used as parallel data. It contains around 2 million aligned sentences from the European Parliament. Table 3 shows the statistics for this corpus.

| Europarl v7 Corpus | English | Spanish |
|---|---|---|
| Number of sentences | 1,965,734 | 1,965,734 |
| Number of tokens | 49,093,806 | 51,575,784 |

Table 3: Statistics of Europarl v7 Corpus

### 3.2 Experiments

We performed a set of experiments in order to test different approaches for aspect-based CLSA. Each experiment requires a different amount of parallel data.

**Representation of Training and Test Data for Sentiment Classification**

For all experiments we use the same train and test split shown in Table 1. For each experiment, we trained a classifier on the English training data, performed the cross-lingual transfer on the Spanish test data and used this new data to test our classifier, as in Figure 1. One difficulty encountered when using vector representations is that the opinion units are variable length. This means that to train a classifier either we find a fixed-length representation for all opinion units or we use a classifier that accepts variable-length input. We decided to take an averaging approach, which has shown promise in other works (Iyyer et al., 2015). For each opinion unit we took the arithmetic mean of the words that compose the opinion unit,

[3]http://www.statmt.org/europarl

as shown in Figure 2, in order to create a fixed-length vector representation for each sentence. We then use these vectors to train a classifier. For the SMT transfer methods, we trained the classifier on unigram features. In all experiments, we used the *sequential minimal optimization* (SMO) classifier from the WEKA toolkit (Hall et al., 2009).

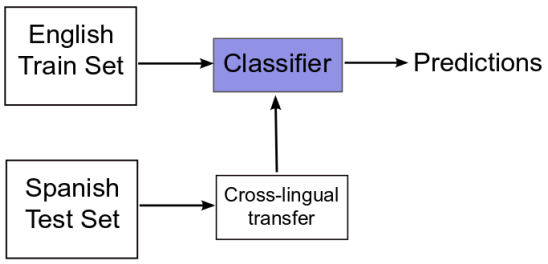


Figure 1: The process of cross-lingual sentiment classification. We assume that the opinion units have already been determined. The English train set is used to train a classifier. The Spanish test set is mapped accordingly and the classifier is tested on this cross-lingual test set.

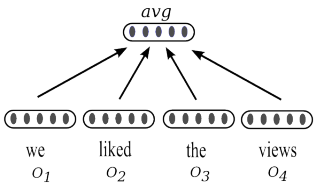


Figure 2: The representation of an opinion unit. For each word $o_n$ in the opinion unit, we take its vector representation and average these vectors in order to create a fixed-length vector $avg = \sum_{i=1}^{n} \frac{o_i}{n}$, which we can use to train our classifier.

**Zero-shot learning**

Since we were interested in using the least amount of parallel data possible, we started with zero-shot learning. This is an approach which attempts to map between two monolingual vector representations using a "translation matrix" *W*. The desired effect is that the dot product of a Spanish word vector and *W* would be similar to the word vector of its English translation, as in Equation 1.

$$\mathbb{R}^{interesante} \circ W \approx \mathbb{R}^{interesting} \tag{1}$$

One could then perform a search for the most similar English vector and use this as a feature for classification. The only parallel data necessary is a bilingual dictionary. Given a pair of translated words and their associated vector representations $\{x_{eng}, x_{spa}\}$, we minimize the cost function in Equation 2 for our training vocabulary of length *n*:

$$\min_{W} \sum_{i=1}^{n} \parallel x_{eng} - x_{spa} \circ W \parallel^2 \tag{2}$$

Following Mikolov et al. (2013b) we created two sets of monolingual word embeddings using the Europarl v7 corpus (Koehn, 2005). We used the Skip-gram model (Mikolov et al., 2013a) and created 300 dimensional vectors using a window of 5 words, and 10 negative samples. We compiled a bilingual dictionary by taking the 8000 most common words in the English Wikipedia and translating them using Bing Translator[4]. Although Bing gives several options, we take only the first translation for use in our

[4]http://www.microsofttranslator.com/

bilingual dictionary. We then removed errors and ambiguous words and arrived at a final number of 4518 word pairs to train the matrix. Finally, we used stochastic gradient descent to optimize the translation matrix *W*. After creating the transition matrix *W*, we tested the effectiveness of this matrix translation to enable CLSC.

For each opinion unit in the corpora, we created a fixed-length vector representation, as shown in Figure 2. We now had a dataset with training instances such as $\{x_i : y_i\}$, where $x_i$ was a 300 dimension vector and $y_i$ was its corresponding label (Strong Positive, Positive, Negative, Strong Negative).

As a baseline, we trained and tested an SVM on the Spanish data from the OpeNER corpus as the Spanish test set has the same opinion units as the cross-lingual test set. We did the same with the English data, although this is not truly comparable[5]. Results are shown in Table 4.

We then created the cross-lingual test set by applying our translation matrix *W* to the Spanish test set. In order to find the most similar vector from the English word embeddings, we used a *k nearest neighbor* algorithm with cosine as the distance metric. Finally, we used the mean of the word embeddings as mentioned above to create the final fixed-length representation. We tested on the cross-lingual test set. The results are shown in Table 4.

From the results it is clear that our zero-shot approach did not yield any effective results. This may be a result of several factors. Mikolov et al. (2013b) were able to leverage a simple mapping strategy between word embeddings created from large monolingual datasets in order to fill the gaps in translation dictionaries. Given the poor results of this experiment, it seems unlikely that this same strategy can effectively capture the complex relationship between target and source language word vectors in a way that is useful for sentiment analysis. This is likely due to the different purposes of the mapping strategy in each approach. In Mikolov et al. (2013b), the success of this technique depended largely on using a small subset of the vocabulary and pairing it with other approaches. In our approach, however, all of the weight of correctly classifying a phrase fell on the accuracy of the mapping scheme. Therefore, it seems that any error in the mapping resulted in the propagation of error during classification.

Another problem that arose is that there were some words whose vector representation always appeared as the nearest neighbor of many other words, although they were not semantically similar with any of them. This problem is known as *hubness* and is an intrinsic problem with high-dimensional vector space. Our work seems to confirm the research of Lazaridou et al. (2015) and Georgiana Dinu et al. (2015), who showed that hubness is compounded when trying to create a linear mapping between two sets of word embeddings.

**Bilingual Word Embeddings**

The next set of experiments required the use of parallel sentences to create bilingual word embeddings (**BWEs**). Following the work of Luong et al. (2015), we created bilingual word embeddings using the Bilingual Skip-gram algorithm, which uses the Skip-gram model (Mikolov et al., 2013a) with an added bilingual objective. This algorithm creates vector representations in which words that appear in parallel sentences have similar representations. We used the Bilingual Skip-gram algorithm to train English and Spanish word vectors on the Europarl corpus (Koehn, 2005) and the corpus of parallel sentences in the hotel domain used in the work of Lambert (2015). We created the alignment using 3 iterations of the Berkley Aligner[6]. We then created 300 dimensional vectors with a window of 5 words, 10 negative samples and ran the algorithm for 3 epochs. This process gave us two sets of word embeddings in which words that often appear in parallel sentences have similar vector representations.

To train our classifier, we used our learned English embeddings and take the average of the vectors in each opinion unit in the English train set. We performed the same procedure with the learned Spanish embeddings and the Spanish test set. The results are shown in Table 4.

The results given by the bilingual word embeddings are not optimal, but are promising enough to warrant more research. There are problems with bilingual word embeddings which would need to be addressed in order to improve their usefulness for CLSC. First, there is the problem of ambiguity that affects all word embeddings. One way to correct this problem would be to disambiguate the word

[5]The monolingual English test set does not have the same examples as the Spanish one.

[6]https://code.google.com/archive/p/berkleyaligner/

senses prior to creating the word embeddings. Cheng et al. (2014) show that this technique improves the performance of distributional models for learning compositional models of meaning and it may improve the performance for sentiment analysis as well.

Secondly, due to the fact that they have similar distributions, antonyms are often given similar vector representations. This is not a problem for POS-taggers or parsers, but it is detrimental to sentiment analysis systems based on word embeddings because these words have opposing polarities and should therefore have different vector representations. To remedy this, one could add a classification task in the problem formulation that would better separate these antonyms into differing vector spaces (Tang et al., 2014). Another option is to decompose the word vectors into interpretable subspaces, train them to differentiate for a certain property, and use only these spaces as features (Rothe and Schütze, 2016).

**Stacked Denoising Autoencoders**

Following the work of Zhou et al. (2016) we trained a stacked bilingual denoising autoencoder (**SDBA**) on parallel sentences from the Europarl corpus. This approach aims to encode the parallel sentences into a common latent space. Given a vocabulary of length $n$, the autoencoder maps the sentences, which are represented as n-dimensional one-hot vectors, to a lower dimensional representation. These representations are then used to reconstruct the original sentences. In order to keep the autoencoder from simply learning the identity function, the lower dimensional representation of one of the sentences is corrupted, which causes the autoencoder to look for discriminative features to help reconstruct the original sentences. In this way, the autoencoder learns to find a lower dimensional representation that encodes as much information as possible needed to reconstruct the bilingual sentences.

We created source and target language autoencoders with 1000 hidden units, which were then mapped to 500 hidden units. The corruption level was set to 0.5. The 500 source and 500 target hidden units were then concatenated and normalized to unit length and fed to the bilingual autoencoder, again with the corruption level set to 0.5. After the autoencoder had been trained, we could use the learned weights to force any of our data into a latent bilingual space.

We then created our training data by mapping the opinion units from the English train set to this lower dimensional latent representation, which was a 500 dimensional vector. We trained a classifier on the mapped English training set. We tested on the similarly mapped Spanish test set. The results are shown in Table 4.

The stacked denoising autoencoder approach gave reasonably good results, despite the fact that it was designed for sentence-level CLSA. There are still ways which we could adapt this approach to aspect-based CLSA. By using word alignment, we could split sentences into parallel or pseudo-parallel n-grams and train the autoencoder with this data. This may improve its performance at aspect-level.

**Statistical Machine Translation**

For the final experiment we used statistical machine translation as a means of bridging the gap between languages. We compared Google Translate[7], a highly developed SMT system, as well as Constrained SMT (see section 2.2). First, we trained our monolingual English classifier and used Google Translate to create our cross-lingual test set. In this case, we translated only the opinionated phrases. This technique has the disadvantage that translation is done without context. We compared this with the Constrained SMT approach in Lambert (2015). This technique allows us to translate the opinion units in context, but without reordering or scrambling them. The language model used in this approach was trained with hotel domain data. All of this improved the quality of translation and resulted in more accurate CLSC results.

Finally, we trained our classifier on unigram features from the monolingual English training set. We created test sets by translating the Spanish test data with each SMT system. The results are shown in Table 4.

The constrained SMT approach is the most accurate approach and shows that, given a more refined treatment of less parallel data, one can achieve CLSA systems which are comparable to monolingual ones. It is interesting that Google Translate has a better BLEU score than constrained SMT[8], but the

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

[8] Google Translate scores a 48.6 BLEU in English-Spanish true-case, versus 45.3 for constrained SMT.

performance on the classification task was lower. It also showed poorer results than the bilingual stacked denoising autoencoder.

**Equal amounts of parallel data**

Each of the previous experiments rely on different amounts of parallel data for optimal performance. Since we are interested in their performance on under-resourced languages, we ran all experiments again with the minimal amount of parallel data (measured at 15.9M English words). The results shown in Table 4 show that, despite a general decrease in precision, recall and F1, the performance of bilingual word embeddings remains stable with less data. The stacked denoising autoencoder, however, performs poorly with this amount of data.

## 4 Results

| | English | Spanish | Zero-shot | Const. SMT | BWEs | SBDA | Google SMT |
|---|---|---|---|---|---|---|---|
| Parallel Data | - | - | 4518 | 15.9M | 15.9M | 15.9M | - |
| Precision | - | - | .453 | .779 | .49 | .254 | - |
| Recall | - | - | .310 | .758 | .468 | .4 | - |
| F1 Score | - | - | .351 | .755 | .473 | .338 | - |
| Accuracy | - | - | 48% | 75.76% | 62% | 55% | - |
| Parallel Data | 0 | 0 | 4518 | 15.9M | 49M | 49M | ? |
| Precision | .824 | .803 | .517 | .779 | .632 | .682 | .670 |
| Recall | .822 | .809 | .503 | .758 | .598 | .590 | .568 |
| F1 Score | .820 | .800 | .434 | .755 | .567 | .633 | .615 |
| Accuracy | 82.22% | 80.86% | 50.25% | 75.76% | 59.76% | 74.5% | 72.81% |

Table 4: Results of Crosslingual Experiments: Precision, recall and F1 are the weighted averages of all classes. The amount of parallel data is measure in the number of English tokens used in training the transfer method.

## 5 Discussion

**Role of Parallel Data**

It is interesting to see that there is not a direct correlation between the amount of parallel data used and the results. Constrained SMT uses less data than bilingual word embeddings or stacked denoising autoencoders and still outperforms both. However, this approach uses higher quality, in-domain data as well as tuning parameters which adapt it to this domain. The trend within representation and distributional approaches has been to use larger and larger datasets , but these results seem to suggest that using smaller, task-specific in-domain datasets which are automatically discovered from larger datasets may be key in improving performance in CLSC.

**Representation**

Besides using the average of the vectors in the opinion unit as a representation, we also experimented with summation and using Long Short-term Memory networks (LSTMs) as a way to deal with the different lengths of the opinion units for our vector-based methods. Although it lacks a strong theoretical motivation, summation is often used in distributional semantics as a baseline for combining vectors (Giorgiana Dinu and Baroni, 2014). Summation led to results that were slightly worse than averaging. This is likely due to the fact that longer opinion units result in vectors which are a magnitude larger than shorter opinion units. We discuss LSTMs below.

**Classifiers**

Apart from the SVM classifiers used in all experiments, we conducted further experiments using deep feed-forward networks during the zero-shot and bilingual word embeddings experiments. We used the

DAN model (Iyyer et al., 2015), with three hidden layers of 300 dimensions. This model performed better on the zero-shot learning experiments, but similar to SVMs on all other experiments.

We also experimented with an LSTM with a 400 dimensional hidden layer. The final state of the LSTM is used to output a softmax probability over the four classes. The results, however, were not competitive with the SVM for zero-shot learning or bilingual word embeddings. We believe the loss of information during transfer did not allow the LSTM to detect the same features during testing that it found while training. We also suspect that the dataset was not large enough to train an LSTM easily. This may limit the usefulness of LSTMs in cross-lingual settings where the dataset used to train the model is small. Given larger training sets, this effect may decrease.

## 6 Conclusion and Future Work

We have presented a comparison of aspect-based CLSA approaches using different amounts of parallel data. The results show that a simple zero-shot learning approach is currently ineffective for CLSA. We show that distributional vector representations are more promising and produce results that are comparable to simple SMT baselines, but still require more research.

In future work, we plan to investigate the role of prior disambiguation and ways to add sentiment-specific information to bilingual word embeddings. We believe this will make bilingual word embeddings more useful for aspect-based CLSA. Another approach that could improve the performance is to use ensemble classification, where we combine SMT and word vector information. Finally, we will extend these techniques to Catalan and Basque.

## References


Agerri, Rodrigo, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, 51(September):215-218.

Balahur, Alexandra and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56-75.

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

Banea, Carmen, Rada Mihalcea, Janyce Wiebe. 2013. Porting multilingual subjectivity analysis resources across languages. *IEEE Transactions on Affective Computing*, 4(2):211-225.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137-1155.

Chandar, Sarath, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. *Advances in Neural Information Processing Systems*, 27:1853-1861.

Cheng, Jianpeng, Dimitri Kartsaklis, and Edward Grefenstette. 2014. Investigating the role of prior disambiguation in deep-learning compositional models of meaning. In *Learning Semantics Workshop NIPS 2014*, 2(1):1-5.

Demirtas, Erkin and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the International Workshop on Issues of Sentiment Discovery and Opinion Mining -WISDOM '13*, pages 9:1-9:8.

Dinu, Georgiana and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* , pages 90:99

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 2015 International Conference on Learning Representations*

Duh, Kevin, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers* , 2:429-433.

Gouws, Stephan, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning* , pages 748-756.

Guo, Yuhong and Min Xiao. 2012. Cross language text classification via multi-view subspace learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* , pages 1615–1622.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10-18.

Hass, Michael and Yannick Versley. 2015. Subsentential sentiment on a shoestring: A crosslingual analysis of compositional classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 694-704.

Hu, Minqing and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168-177.

Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daume III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681-1691.

Kiritchenko, Svetlana, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 437-442.

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. 2012. Inducing cross-lingual distributed representations of words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Technical Papers*, pages 1459-1474.

Klinger, Roman and Phillip Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the 19th Conference on computational Natural Language Learning*, pages 153-163.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit* .

Lambert, Patrik. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 781-787.

Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 781-787.

Lin, Zheng, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. 2014. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CiKM '14*, pages 1089-1098.

Lu, Bin, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 320-330.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. *NAACL Workshop on Vector Space Modeling for NLP.*

Mihalcea, Rada, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 976-983.

Mikolov, Tomas, Greg Corraddo, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representation in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1-12.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. In arXiv: 1309.4168.

Moghaddam, Samaneh. 2010. Opinion Digger: An unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, pages 1825-1828.

Mohammad, Saif M., Mohammad Salameh, and Svetlana Kiritchenko. 2015. How translation alters sentiment. *Journal of Artificial Intelligence Research*, Volume 55, pages 95-130.

Pan, Junfeng, Gui-Rong Xue, Yong Yu, and Yang Wang. 2011. Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 289-300.

Prettenhofer, Peter and Benno Stein. 2011. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology*, 3(1):95-130.

Rothe, Sacha and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspace. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* , pages 512:517.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embeddings for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555-1565.

Vulić, Ivan and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, Volume 55, pages 953-994.

Wan, Xiaojun. 2009. 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 Conference on Natural Language Processing*, pages 235-243.

Zhou, Guangyou, Zhiyuan Zhu, Tingting He, and Xiaohua Tony Hu. 2016. Cross-lingual sentiment analysis with stacked autoencoders. *Knowledge and Information Systems*, 47(1):27-44.

Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao. 2012. Cross-language opinion target extraction in review texts. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 1200-1205.

Zhu, Jinbo, Huizhen Wang, Benjamin Tsou, and Muhua Zhu. 2009. Multi-aspect opinion polling from textual reviews. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1799-1802.