UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LUCAS RAFAEL COSTELLA PESSUTTO

Clustering Multilingual Aspect Phrases for Sentiment Analysis

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. Viviane Pereira Moreira

Porto Alegre
January 2019

CIP — CATALOGING-IN-PUBLICATION

Pessutto, Lucas Rafael Costella

Clustering Multilingual Aspect Phrases for Sentiment Analysis / Lucas Rafael Costella Pessutto. – Porto Alegre: PPGC da UFRGS, 2019.

69 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Advisor: Viviane Pereira Moreira.

1. Aspect-based sentiment analysis. 2. Multilingual aspect clustering. 3. Unsupervised learning. 4. Word embeddings. I. Moreira, Viviane Pereira. II. Título.

“Imagination is more important than knowledge.
Knowledge is limited. Imagination encircles the world.”
— Albert Einstein

AGRADECIMENTOS

Agradeço aos meus pais Sérgio e Tania, ao meu irmão Leonardo, a minha família e todos os meus amigos. Obrigado pelo incentivo, paciência e por estarem do meu lado durante toda a minha jornada.

Agradeço a minha orientadora, Prof. Viviane Moreira, que durante este trabalho contribuiu com seu vasto conhecimento e experiência. Mais do que isto, seu incentivo foi indispensável em diversos momentos durante a execução deste trabalho.

Dedico este trabalho a todos os professores que participaram da minha formação, desde os que me ensinaram a ler e escrever, aos que me ensinaram a programar, e principalmente aos que me fizeram gostar de ciência. Graças a profissionais tão brilhantes este trabalho pode se tornar realidade.

Agradeço aos meus colegas de curso, por deixarem mais fácil e mais leve a minha caminhada e pelas trocas de conhecimento e experiência que tivemos. Agradeço também à toda a comunidade do INF-UFRGS que se empenham pra garantir uma formação de excelência aos seus estudantes.

E finalmente agradeço a Deus por ter proporcionado estes agradecimentos a todos que tornaram minha vida melhor, além de ter me dado uma família maravilhosa e amigos sinceros.

Lucas Rafael Costella Pessutto

Agrupamento de Expressões de Aspecto Multilíngues em Análise de Sentimentos

RESUMO

A pesquisa em análise de sentimentos obteve um significante desenvolvimento nos últimos anos motivado pela crescente disponibilidade de comentários opinativos sobre produtos. Mais especificamente, tem havido um crescente interesse em análise de sentimentos baseada em aspectos, cujo objetivo principal consiste em extrair, agrupar e avaliar a opinião global em relação às características da entidade que está sendo avaliada. As técnicas existentes para extração de aspectos podem produzir uma quantidade excessiva de aspectos – muitos destes relacionados a uma mesma característica do produto. Este problema é agravado quando os comentários estão escritos em muitos idiomas. Este trabalho aborda a tarefa de agrupamento de aspectos multilíngues, que consiste em criar grupos de aspectos semanticamente relacionados, extraídos de comentários escritos em diversos idiomas. Este trabalho propõe uma técnica não supervisionada para esta tarefa. Ela baseia-se na informação contextual advinda dos aspectos, que é representada através de *word embeddings*. Esta representação aliada a uma medida de similaridade (Word Mover's Distance) permitiu realizar o agrupamento de aspectos relacionados, utilizando o algoritmo k-means. A contribuição deste trabalho inclui as técnicas para resolver este problema juntamente com os testes realizados em comentários escritos em cinco idiomas. Os experimentos mostraram que a técnica não supervisionada de agrupamento alcança resultados que superam um *baseline* semi-supervisionado.

Palavras-chave: análise de sentimentos baseada em aspectos, agrupamento de aspectos multilíngue, aprendizagem não supervisionada, word embeddings.

ABSTRACT

The area of sentiment analysis has experienced significant developments in the last few years. More specifically, there has been growing interest in aspect-based sentiment analysis in which the goal is to extract, group, and rate the overall opinion about the features of the entity being evaluated. Techniques for aspect extraction can produce an undesirably large number of aspects – with many of those relating to the same product feature. This problem is aggravated when the reviews are written in many languages. We address the novel task of multilingual aspect clustering which aims at grouping together semantic related aspects extracted from reviews written in several languages. Our method is unsupervised. We rely on the contextual information of the aspects, which was represented through word embeddings in our approach. This representation allied with a good similarity measure (Word Mover's Distance) allows us to cluster together related aspect phrases, using k-means algorithm. We contribute with a proposal of techniques to tackle this problem and test them on reviews written in five languages. Our experiments show that our unsupervised clustering technique achieves results that outperform a semi-supervised baseline.

Keywords: Aspect-based sentiment analysis. multilingual aspect clustering. unsupervised learning. word embeddings.

LIST OF ABBREVIATIONS AND ACRONYMS

ABSA	Aspect Based Sentiment Analysis
LSA	Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
nBow	Normalized Bag-of-words
WMD	Word Mover's Distance

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The dawn of the Web 2.0 changed the way users interact on the Internet, enabling more content production as people express their opinions over many subjects on multiple platforms. E-commerce systems allow users to give opinions about the products that are sold. This information, in turn, becomes useful to other users as they can rely on previous shopping experiences from other people as a basis for their own purchases. In addition, companies can take advantage of the opinions to measure the acceptance of a product and improve it according to the users' taste.

While useful and valuable, reviews are difficult to process because they are often represented as large amounts of unstructured text. Moreover, in systems accessed on a global scale, opinions can be found in different languages, posing further difficulties to automatic processing.

Sentiment Analysis is the field of study which aims at processing the information conveyed by unstructured texts, providing structured information that facilitates the understanding of the opinions, attitudes, or emotions towards a particular entity (LIU, 2011). The main tasks in sentiment analysis include polarity attribution, aspect extraction, and opinion summarization. Polarity attribution consists in determining if the opinion expressed in a review is positive, negative, or neutral. Aspect Extraction is a more fine-grained task as its goal is to extract the features of the entity to which the opinion is targeted. Opinion summarization aims to build a concise text that synthesizes the opinions about the entity from a large set of review texts (ZHANG; LIU, 2014).

In spite of the good results achieved by modern aspect extraction techniques, they can produce an undesirably large number of aspects. This happens due to the fact that people use different words to express the same aspect of an entity (LIU, 2011). For example, the words *screen*, *display*, and *touchscreen* refer to the same feature in the smartphone domain. In order to group together the terms that refer to the same feature, *aspect clustering* is employed. This is a fundamental step to allow the construction of summaries containing a small list of representative aspects that convey the users' overall opinion.

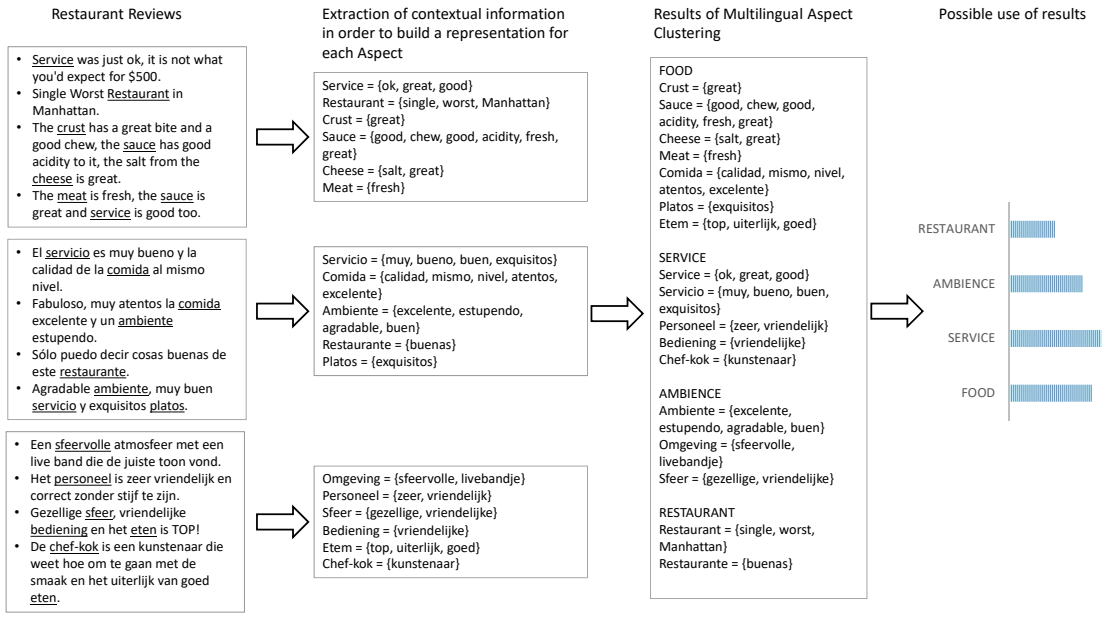
Reviews in multiple languages are abundant in a number of important sources such as TripAdvisor, AirBnB, Amazon *etc.*, as a consequence, dealing with multilingual data becomes necessary. In this scenario, one language will be less represented than others lacking the required amount of data to allow for sentiment-analysis algorithms to yield good results. In such cases, it is useful to rely on languages with more density of re-

views. The combined use of multiple languages for sentiment analysis has proven useful and enabled reaching results that are significantly better than when a single language is considered (BANEA; MIHALCEA; WIEBE, 2010).

The focus of this work is on *multilingual aspect clustering*, which can be defined as the task of grouping together equivalent aspects across multiple languages. To the best of our knowledge, this is the first work to address this problem. Our solution combines unsupervised clustering, syntactic term similarity, and word embeddings. We carried out experiments on restaurant reviews written in English, Spanish, Russian, Dutch, and Turkish and compared our performance against an established baseline. The results show that our unsupervised clustering technique achieves results that are better than the results of the semi-supervised baseline.

An example of Multilingual Aspect Clustering can be seen in Figure 1.1. Initially, the set of reviews is formed by the union of the subsets of reviews in three languages (English, Spanish, and Dutch) with the aspect phrases already extracted. The next step is to build a language independent representation for each aspect phrase. Then, we can apply a clustering algorithm on the aspects' representations in order to group together semantically similar aspect phrases. In our example, four clusters are formed representing the aspects Food, Service, Ambience, and Restaurant. Once the data is clustered, we can use it to build summaries that synthesize the sentiment expressed by the reviewers.

Figure 1.1: An example of Multilingual Aspect Clustering



Source: The author

Structure of this work

The remainder of this document is structured as follows. The next chapter presents background knowledge that supports our multilingual aspect clustering algorithm. In Chapter 3, the related work is discussed. Three fields are covered in our review – sentiment analysis, focusing on aspect extraction; monolingual aspect clustering; and multilingual document clustering. Once the basic concepts have been presented, in Chapter 4, our approach to the Multilingual Aspect Clustering problem is presented. Chapter 5 shows the experiments, results obtained, and presents a discussion over that results. Finally, Chapter 6 concludes this thesis, summarizing our contributions and discussing possibilities of future work.

2 BACKGROUND

This chapter aims to present the knowledge necessary for a better understanding of this work. We start by defining the problem of multilingual aspect clustering. In sequence, Vector Representation of Words, Centroid-Based clustering algorithms, and Distance Measures are discussed.

2.1 Sentiment Analysis

Opinions abound on the Internet nowadays. They are a valuable source of information since people can rely on them before making a purchase, giving support to their decisions. Companies also can benefit from this huge amount of opinions, as they do not need to conduct opinion polls and focus groups to measure acceptance of a particular product (LIU, 2011). The large volume of opinions available becomes hard for a human to process. It leads to the study of ways of automating the processing of opinions, in order to summarize them.

Liu (2012) defines an opinion as a quintuple $O = (e_i, a_{ij}, s_{ijnp}, h_k, t_p)$, where e_i is the entity being evaluated, which can be a product, a service, a topic, an issue, a person, an organization, or an event. a_{ij} is the aspect of the entity being reviewed, s_{ijnp} is the sentiment related to the aspect expressed on the review, h_k represents the opinion holder (person who emits the opinion), and t_p is the time when the opinion was emitted.

Considering Figure 2.1 which contains a review extracted from TripAdvisor, we can highlight the components of an opinion. The entity e_i in this review is the Hotel Holiday Inn NYC. The reviewer expresses opinions about two aspects of the hotel, so we have $a_i = \{\text{rooms, staff}\}$.

The sentiment expressed by the reviewer can be seen in two different levels of granularity in this review. We can analyze the overall feeling about the hotel expressed by the evaluation of four "stars" out of five given by the reviewer. We also can measure the sentiment about each aspect evaluated in the review. The user evaluated two aspects of the hotel. Regarding the aspect rooms, the opinion is that they are clean, but very small. This can be considered a NEUTRAL opinion, because the reviewer expressed a positive and a negative opinion about the hotel's room. Next, the user presents a POSITIVE opinion regarding the staff of the hotel, using the opinion word *wonderful* to describe it. The last opinion components are the opinion holder (h_k) which is the TripAdvisor's user "Paula

Figure 2.1: Components of an opinion

Source: <https://www.tripadvisor.com/>

K", and the time when the opinion was emitted (t_p) that is June 19th, 2018.

As shown in the example, Sentiment Analysis can be performed at different granularity levels. The lowest granularity level is the Document one, whose objective is to classify the reviews in three pre-defined categories – POSITIVE, NEGATIVE or NEUTRAL. At Sentence Level, each sentence in the review is classified in the three categories above (ZHANG; LIU, 2014).

Aspect-Based Sentiment Analysis (ABSA), also known as feature-based opinion mining is a more fine-grained task used to summarize reviews. It aims to extract the features (or aspects) of the entity which is being evaluated. Zhang and Liu (2014) point out the three sub-tasks of ABSA: (i) Identify and extract entities in reviews; (ii) Identify and extract the aspects of an entity; and (iii) Determine the sentiment over the entities and the aspects. For example, in the sentence *"Burger King has the best french fries I ever ate."*, we can identify "*Burger King*" as the entity, "*french fries*" as the aspect, and finally the adjective "*best*" indicates a positive opinion about the aspect.

Sentiment Analysis (especially at aspect level) is a challenging task. Each user expresses their opinion in different ways using free text, which causes a diversity of vocabulary by the introduction of informal language – misspelled words, abbreviations, variations of the same word form, *e.g.,* huge, huuge, or huuuuuuuuuge, use of emoticons and emojis, *e.g.,* :−) , =) , or ☺. In addition, most of the existing techniques for sentiment analysis at aspect level have difficulties in dealing with comparative sentences, *e.g., "The hotel X is better than hotel Y"* or discovering implicit aspects in the reviews. Implicit aspects are aspect words that are not nouns or noun phrases, *e.g., "Hotel K is expensive"*, where the

adjective expensive implies the aspect price (LIU, 2012).

Another challenging aspect regarding Sentiment Analysis is multilingualism. Since in practice reviews can be written in any language, proposed solutions need to analyze datasets containing reviews in two or more languages. This requirement poses further challenges. First of all, most of the existing techniques for Sentiment Analysis are language dependent, which means that the features used by this kind of algorithms rely on specific language resources, such dictionaries and lexicons, or on the assumption that the documents share the same vocabulary. Another problem faced by these approaches is that some languages are poorer in language resources, which are costly to create. On the other hand, studies like Banea, Mihalcea and Wiebe (2010) and Balahur and Perea-Ortega (2015) indicate that the combination of multilingual features tends to produce better results than the monolingual ones.

2.2 Multilingual Aspect Clustering

In the context of multilingual aspect clustering, the set of input reviews can be defined as $R = \{R_1, R_2, ..., R_l\}$, where R_l corresponds to the subset of reviews in language l, with $l >= 2$. All reviews belong to the same domain, for example, if R_l contains opinions about smartphones, all other subsets will also have reviews on smartphones. Each subset is composed of reviews $R_l = \{r_{l_1}, r_{l_2}, ..., r_{l_m}\}$, where r_{l_m} denotes the m^{th} review in language l.

Aspect extraction techniques may be employed on the reviews in order to extract the properties of the target entity (please refer to Section 3.3 for more details on aspect extraction). The explicit properties of a target that occur in a set of reviews are referred to as *aspect phrase* (also called *product feature* or *surface form* in the literature). An aspect phrase is composed of one or more terms (*e.g., battery*, *battery life*). The formal definition of an aspect phrase follows that one presented by Liu (2012) in Section 2.1. The set of aspect phrases (AF) will be the union of all reviews' aspect sets.

Users can express the same product features using different words or phrases. Thus, a clustering step is necessary to group the aspect phrases that belong to the same category. Each of these groups will be called *Aspect Group* and will consist of a set of aspect phrases in multiple languages. We can formally define the problem of multilingual aspect clustering as the mapping of the aspects in the AF set into a AG set where $AG = \{c_1, c_2, ..., c_k\}$. Each subset c_k of this set will contain several aspect phrases referring to the same Aspect

Group. k is the total number of aspect groups.

Two important properties of the AG set should be highlighted. First, the union of all subsets c_k in AG results in the set of aspect phrases AF. It means that all aspects have to be assigned to a group. The second property is that the intersection of the subsets of AG will be the empty set because every aspect phrase belongs to just one aspect group.

2.3 Vector Representation of Words

The goal of representing words in a vector space is to map semantic similarity between them. The techniques employed for this task are based on the hypothesis that words with the same semantic meaning are used in the same contexts along the documents.

In order to express the correlation between words, we can represent them through one-hot vectors, where each word is represented by a vector of zeros and ones, where zero denotes that the two words do not co-occur and one that the words co-occur in documents. The size of each word vector will be equal to the size of the vocabulary, which usually tends to create huge sparse vectors. That makes these vectors difficult to create, process, and store.

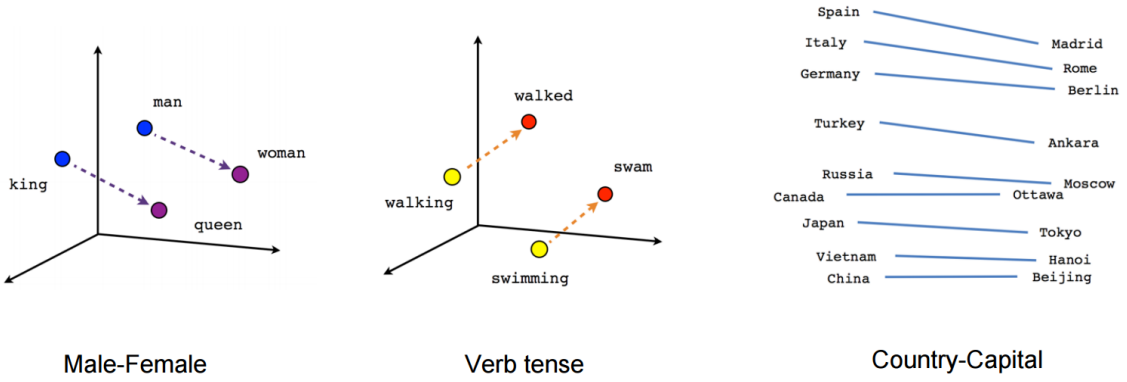
Research aiming the dimensionality reduction of word vectors resulted in the emergence of a new vector representation of words denominated word embeddings. In this language model, the words are represented as low dimensional vectors, keeping the distributional similarity between them (MIKOLOV et al., 2013).

An interesting feature of the word embeddings is the ability to map linguistic regularities present in documents. As can be seen in Figure 2.2, several linguistic regularities such as gender, verbal tense and even the relationship between a country and its capital can be found in word embeddings space.

For example, if we take the vector of the word ‘king’ and subtract the vector of ‘man’ and then add the vector of the word ‘woman’, we will get close to the vector of the word ‘queen’. Once we know this pattern, it is easy to find new word pairs that follow it. This semantic properties of word embeddings can be useful in many Natural Language Processing applications and also can be used to evaluate the quality of the word embeddings itself (MIKOLOV; YIH; ZWEIG, 2013).

Mikolov et al. (2013) proposed the most well-known word embeddings model, *word2vec*, which is an efficient and fast training method for word embeddings. The authors devised two model architectures for the word vectors training – continuous bag of words

Figure 2.2: Examples of semantic relationships between words in a Word Embedding Space



Source: <https://www.tensorflow.org/tutorials/representation/word2vec>

(CBOW) and skip-gram. Both approaches consist of neural networks trained to predict neighbor contextual words.

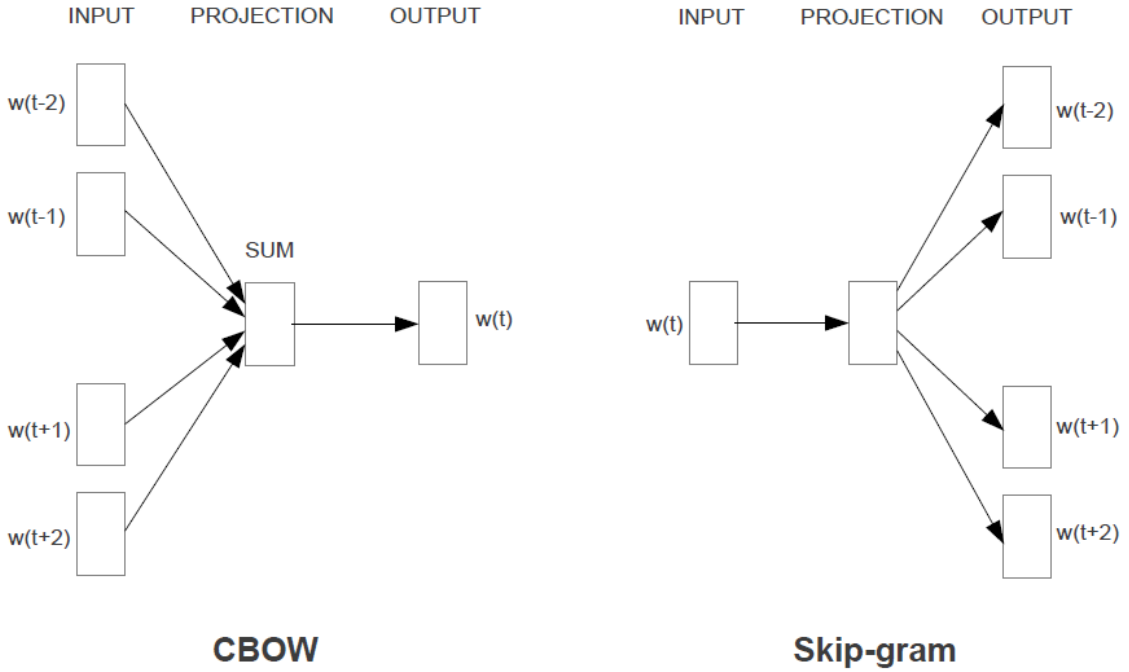
This kind of model tries to assign a probability to a sequence of tokens. The goal is to score high probabilities to sentences that are syntactic and semantically correct. For example, the sentence *"The menu list is extensive"* will have a high probability score, while the sentence *"list sentence the extensive is"* will not. The probability of a word w_t in a sequence of words is given by the conditional probability between that word and the words in his context. We can represent the context words as $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$, where t indicates the position of a word in the text.

In the CBOW model, the context words are used to predict the current word w_t. In the sentence *"The wine list is extensive"*, for example, the CBOW model will try to predict the word "list", having as input the context words ["the", "wine", "is", "extensive"]. Its architecture is shown in Figure 2.3. CBOW receives the context words as an input of a neural network. Next, a projection layer consisting of a softmax function gets the average of the input vectors and computes the conditional probability to export to the output layer. The objective function of the training is to make the generated probabilities be like the true probabilities given by the one-hot vectors.

The skip-gram model works almost like the CBOW. The main difference between the two models is that skip-gram receives as input the center word and tries to predict the words in the context. Figure 2.3 shows the organization of the neural network. Considering the example above, from the word "list" the skip-gram model tries to predict the context words ["the", "wine", "is", "extensive"].

Some applications need to handle texts in multiple languages, which lead to the development of multilingual word embeddings. In this kind of representation, words from

Figure 2.3: CBOW and Skip-gram Models



Source: (MIKOLOV et al., 2013)

different languages share the same vector space. According to Ruder (2017), this kind of word embedding has the advantage of sharing semantics between words in different languages and the capacity of knowledge transfer from languages with rich resources to the ones with scarce resources. These factors motivated the emergence of several approaches in recent times.

Some approaches learn a matrix capable of performing the linear transformation of monolingual word embeddings in language x in the word embeddings of a language y. These approaches usually rely on bilingual dictionaries or other translation tools, in order to create pairs of words from different languages. Many techniques also trained multilingual word embeddings from parallel or comparable multilingual sources (RUDER, 2017).

In this work, we take advantage of the semantic power provided by word embeddings in order to represent the aspect phrases and the contextual information. We employ multilingual word embeddings in that representation, which allows us to represent aspect phrases in different languages in the same vector space.

2.4 Centroid-Based Clustering Algorithms

Clustering algorithms aim to group data points so that data within the same cluster have high similarity compared with data belonging to different clusters (TAN et al., 2013). This is an unsupervised technique once it does not need to know the class of the data points beforehand.

Centroid-Based Clustering Algorithms (also known as center-based clustering algorithms) assume that data points are distributed in a Euclidean space, so each cluster can be represented by its centroid, which usually is the average of the cluster data points (LESKOVEC; RAJARAMAN; ULLMAN, 2014). The first clustering algorithm based on centroids was k-means. This algorithm is the most widely used for unsupervised clustering (TAN et al., 2013). We will focus the rest of this section explaining this technique.

Algorithm 1 shows the pseudocode for the k-means algorithm. The number of clusters (k) is the only hyperparameter of k-means. The first step of the algorithm is to choose k initial cluster centroids. The simplest way to do this is randomly select data points as the centroids. However, a bad initial cluster set can produce bad result clusters. To work around this problem, one can employ some heuristics to select better centroids. For example, select data points that are as far away as another as possible. However, such heuristics do not guarantee better clusters and also can be hard to compute. The heuristic cited above is an example of that fact. For large datasets, it is necessary to compute the similarity between all data points to select the further one, which is computationally hard to do. Also, further data points can indicate outliers, which would lead the algorithm to poor results (TAN et al., 2013).

Another technique adopted to deal with centroid selection is to run the k-means algorithm many times, varying the initial centroids and reporting as the best cluster the one that achieves the best results (minimum sum of squared errors) (ZAKI; JR., 2014).

Once the centroids have been selected, the remaining data points are assigned to a cluster. For that, the similarity between the data points and each one of the centroids is computed. The data points will be assigned to the most similar cluster (the one with the highest similarity between the data point and the cluster's centroid). After that, new cluster centroids are calculated, based on the average of data points belonging to that cluster. Then, the data points are reassigned to the clusters based on the new centroids. This procedure repeats until convergence, that is when the points do not change clusters anymore.

A variation of the k-means algorithm was proposed in order to cluster data points

ALGORITHM 1: k-means Algorithm. Adapted from (TAN et al., 2013).

```
Input  : k – Number of Clusters
         points – Data Points
Output : C = {c_1, c_2, ..., c_k} – Set of Clusters
:
1 Select k points as initial centroids of each cluster in C
2 repeat
3 |   for point in points do
4 |   |   Assign point to its closest centroid
5 |   end
6 |   Recompute the centroid of each cluster C
7 until points do not change cluster
```

when the number of clusters k is unknown. This approach is called *Bisecting k-means*. Instead of giving the number of desired clusters as input, the user informs a threshold, which represents the maximum number of elements allowed in each cluster.

We can see a representation of Bisecting k-means in Algorithm 2. In the initialization step, all data are assigned to the same cluster (Lines 1–3). Next, the algorithm starts a loop until it reaches the stopping condition (Lines 4–8). This loop consists of two phases, the selection step, in which the cluster with most elements is selected (Line 5) , and the division step, in which the selected cluster is bisected through the application of the k-means algorithm with $k = 2$ (Line 6) (BAEZA-YATES; RIBEIRO-NETO, 2011).

ALGORITHM 2: Bisecting k-means. Adapted from (BAEZA-YATES; RIBEIRO-NETO, 2011).

```
Input  : s – Threshold
         points – Data Points
Output : C = {c_1, c_2, ..., c_k} – Set of Clusters
:
1 for each data_point in points do
2 |   data_point.cluster ← 0
3 end
4 repeat
5 |   data ← max_cluster(points)
6 |   clusters ← kmeans(k = 2, points = data)
7 |   update_data(points, clusters)
8 until length of all clusters < s
```

We take advantage of centroid-based clustering algorithms, more specifically k-means and Bisecting k-means in our approach. These algorithms were chosen because they are well known, widely used and do not require the comparison between all data points, which is a costly task.

2.5 Distance Measure

A key aspect concerning clustering algorithms is the choice of the similarity measure. This is crucial for the good performance of this class of techniques. In addition to correctly expressing the distance between the data points, a similarity measure needs to guarantee the convergence of the k-means algorithm. Euclidean, Manhattan, Jaccard, and Cosine distances are widely used for this purpose. The first two measures are applied in Euclidean Spaces while Jaccard and Cosine are more suitable for documents.

Next, we present the Word Mover's Distance (WMD), which is a measure more suitable when the data points are documents represented as a set of Word Embeddings. It was proposed by Kusner et al. (2015) as a special case of the Earth's Mover Distance, which measures the distance between two probability distributions over a region and is a well-known distance metric. WMD takes advantage of the semantic relationships present in word embeddings.

The goal of WMD is to measure the minimum traveling cost of the word embeddings in one document to the other document. The WMD between two documents is calculated by the summation of the smallest distance between each word in the first document and the words in the second one. It works well even if the two documents have no words in common. For example, the sentences "*The wine list has interesting good values*" and "*They have a good beverage menu with reasonable prices*" has a small WMD score as they contain words that are semantically related *e.g., wine* and *beverage*, *list* and *menu*, and *values* and *prices*.

To compute WMD, it is necessary to represent the documents as normalized bag-of-words (nBoW). This representation is similar to the traditional bag-of-words model, with the addition of weights for the words. Consider that a word i appears c_i times in a document. We can compute its weight (d_i) as shown in Equation 2.1.

$$d_i = \frac{c_i}{\sum_{j=1}^{n} c_j} \tag{2.1}$$

For example, in the sentence "*The wine list has interesting good values*", after removing the stopwords, we have the following bag-of-words representation: [wine, list, interesting, good, values]. Each word appears once in the document, which makes the weight of every word as $1/5$.

The similarity between word pairs is incorporated into the WMD computation by the Euclidean distance between the word embedding representations, $c(i, j) = ||x_i - x_j||_2$,

where x_i is the word embedding for word x. To calculate the WMD between two documents d and d', it is necessary to build a flow matrix (T), where $T_{i,j}$ will represent how much of a word i in d travels to word j in d'. Two conditions need to be satisfied to ensure the transformation of entire document d into the document d' – $(i) \sum_j T_{i,j} = d_i$ and $(ii) \sum_i T_{i,j} = d_j$.

The distance of document d to document d' will be the minimum cumulative cost required to transform all words in d into the words in d_i. Equation 2.2 represents that cost.

$$WMD(d, d') = \sum_{i,j} T_{i,j} c(i, j) \tag{2.2}$$

For example, let d = *"The wine list has interesting good values"* and d' = *"They have a good beverage menu with reasonable prices"*. As shown above, the weight of the words in d are $1/5$. Removing the stopwords in d', we have the following nBOW representation [good, beverage, menu, reasonable, prices], which leads word weight of $1/5$. In Figure 2.4 we compute the Euclidean distance between the the words in d and d'.

Figure 2.4: Example 1 – Euclidean Distance between word embeddings

good	beverage	menu	reasonable	prices	
1.27	0.87	1.15	1.30	1.18	wine
1.31	1.33	1.26	1.28	1.31	list
0.96	1.28	1.24	1.04	1.27	interesting
0.00	1.29	1.27	0.93	1.27	good
1.22	1.30	1.26	1.21	1.14	values

Source: The author

We can now build matrix T and compute the WMD score. For that, we have to choose word pairs with minimum transport cost. Since the size of the two documents is the same, this can be done by matching the words with smallest Euclidean distance values, like 'wine' and 'beverage', and 'prices' and 'values'. Figure 2.5 shows the result Matrix T.

Figure 2.5: Example 1 – Matrix T of weights

$T =$

good	beverage	menu	reasonable	prices	
0	1/5	0	0	0	wine
0	0	1/5	0	0	list
0	0	0	1/5	0	interesting
1/5	0	0	0	0	good
0	0	0	0	1/5	values

Source: The author

Note that if we sum the lines and columns in matrix T, we obtain the weights of the words in the documents, which means the two conditions of WMD computation were satisfied. The WMD of d and d' is computed according to Equation 2.2, as shown in Equation 2.3.

$$\begin{aligned} WMD(d, d') &= 1/5 * 0 + 1/5 * 0.87 + 1/5 * 1.26 + 1/5 * 1.04 + 1/5 * 1.14 \\ &= 0.862 \end{aligned} \tag{2.3}$$

When we have documents of different lengths, the words in the smallest one pairs up with multiple words in the other document, splitting their weights. Let us see another example, considering d = *"The wine list has interesting good values"* and d' = *"Best beer menu"*. The weights of words in d is $1/5$, while in d' is $1/3$. Figure 2.6 shows euclidean distance between words in d and d'.

Figure 2.6: Example 2 – Euclidean Distance between word embeddings

	best	beer	menu	
⎡	1.28	0.87	1.14	⎤ wine
⎢	1.27	1.33	1.26	⎥ list
⎢	1.20	1.29	1.24	⎥ interesting
⎢	1.08	1.28	1.27	⎥ good
⎣	1.29	1.32	1.26	⎦ values

Source: The author

Figure 2.7 shows the matrix T obtained after the pairing of the words in the two documents. Note that the most similar words in the documents, like *good* and *best*, and *beer* and *wine* are paired together and transfer all their weight between each other. The remaining words (*interesting* and *values*) had to split their weight in order to obey the conditions of WMD computation.

Figure 2.7: Example 2 – Matrix T of weights

$$T = \begin{array}{c} \begin{matrix} \text{best} & \text{beer} & \text{menu} \end{matrix} \\ \begin{bmatrix} 0 & 1/5 & 0 \\ 0 & 0 & 1/5 \\ 2/15 & 1/15 & 0 \\ 1/5 & 0 & 0 \\ 0 & 1/15 & 1/5 \end{bmatrix} \end{array} \begin{matrix} \\ \text{wine} \\ \text{list} \\ \text{interesting} \\ \text{good} \\ \text{values} \end{matrix}$$

Source: The author

Once we have the distance between words and the matrix T are calculated, we can

compute the WMD between the documents, as shown in Equation 2.4

$$\begin{aligned} WMD(d, d') &= 1/5 * 0.87 + 1/5 * 1.26 + 1/5 * 1.08 + 2/15 * 1.2 + 1/15 * 1.29 + 2/15 * 1.26 \\ &+ 1/15 * 1.32 \\ &= 1.144 \end{aligned} \tag{2.4}$$

Comparing the results in two examples, in Equations 2.3 and 2.4, we can see that the first pair of documents is more similar than the second one because the first group of documents has more words related between themselves than the second one. The more non-related words two documents have, the greater will be WMD score of these documents.

WMD was used in this work because of its ability to quantify the existing semantics between two documents. This measure is advantageous when dealing with multilingual documents, since it has good results even if two documents have completely different words, which is common in multilingual documents.

2.6 Summary

This chapter addressed the theoretical background necessary for a better understanding of this work. Initially, it presented a formal representation of the Multilingual Aspect Clustering problem. Next, it showed how the aspect phrases and documents can be represented, clustered and how to evaluate distances between these elements. Chapter 4 explains how we use these concepts to address the MAC problem.

The next chapter presents the state-of-the-art in the field of aspect clustering in Sentiment Analysis and Multilingual Document Clustering, which contextualizes the contribution of this work.

3 RELATED WORK

In this chapter, we review the related literature about two key topics: Sentiment Analysis (emphasizing multilinguism, aspect extraction task, and aspect clustering) and Multilingual Document Clustering.

3.1 Sentiment Analysis

A few years ago, Feldman (2013) mentioned that over 7000 research papers had been written about Sentiment Analysis. This demonstrates the significant interest in this area, which aims at labeling texts of different granularities (entire reviews, sentences, and aspects) and different levels of analysis.

The most common approach in document and aspect levels is polarity classification. That technique aims to classify the sentiment express by the reviews in predefined classes, usually positive, neutral and negative. According to Ravi and Ravi (2015), the techniques used to solve this problem can be grouped into three categories Machine Learning Based, Lexicon Based, and Hybrid Approaches.

Machine Learning Based Approaches: This category of approaches consists of extract some features from the reviews and apply supervised or unsupervised algorithms in order to determine the polarity of the reviews. In recent years the use of deep learning techniques was also used in this task. The most common features used in machine learning techniques are terms and n-grams and their frequencies, part of speech tagging, sentiment lexicons, syntactic dependency between words, and sentiment shifters (words that can change the orientation of a sentiment, like the word "not" in the sentence *"This cell phone is not good"*) (YUE et al., 2018).

The techniques employed in order to classify the sentiment of text/sentences in recent works include the use of all kinds of machine learning algorithms. For example, we can cite the use of **supervised algorithms**, like Support Vector Machines and Regression Models (ERTUGRUL; ONAL; ACARTURK, 2017), Rule-Based Approaches (ASGHAR et al., 2017) and Hidden Markov Models (KANG; AHN; LEE, 2018). In the **unsupervised algorithms** highlights the use of k-means algorithm (RIAZ et al., 2017) and Fuzzy C-Means (PHU et al., 2017). Trying to taking advantage of the benefits of supervised and unsupervised approaches, the **semi-supervised algorithms** are employed in sentiment classification task. Self-learning, co-training (IOSIFIDIS; NTOUTSI, 2017)

and topic models (XIANG; ZHOU, 2014) are among the most used techniques. Finally, **deep learning approaches** gained attention last years in this task. LSTM with attention models (CHEN et al., 2016), deep memory networks (DOU, 2017), autoencoders (ZHAI; ZHANG, 2016) and convolutional neural networks (JU; YU, 2018) are between the techniques used in sentiment classification task.

Lexicon-Based Approaches: This category of approaches relies on pre-built resources, like lexicons and dictionaries, containing words expressing some sentiment together with their polarity. The main idea of this approach is to estimate the sentiment score of a document/sentence (GIACHANOU; CRESTANI, 2016). The advantage of lexicon-based techniques is that they do not need training data. On the other hand, build this kind of resource is costly and requires manual annotation. This kind of resource is not good in dealing with domain specific terms, for example, the word "long" changes its meaning depending on the context. It can have a positive orientation, like in *"My notebook has a long battery life"*, or a negative one as in *"We wait for so long to have your order taken by the waiter"* (ISMAIL; BELKHOUCHE; ZAKI, 2018). For that reason, some techniques can create a lexicon for an initial set of seed words. Li et al. (2018), for example, propose a technique to find new sentiment words and an expansion method in order to create a richer domain-specific lexicon.

Saif et al. (2016) proposed SentiCircles, a word representation technique, which can capture the contextual semantic and sentiment of words, based in their co-occurrence pattern. The terms in SentiCircles are represented as a 2D geometric circle, and the sentiment orientation is obtained by using trigonometric identities over that representation. In their work, Khan, Qamar and Bashir (2017) propose a dictionary-based approach for Sentiment Analysis. They use SentiWordNet in order to measure the polarity score, and improve this score by the computation of the sentiment strength, which is a value between -1 and 1. Chi-Square, GSS Coefficient, Odds Ratio and Expected Likelihood Estimate are employed to measure the strength of each entry.

Hybrid Approaches: This set of techniques combines machine learning algorithms with lexicon resources in order to take the advantages of both approaches. For example, Asghar et al. (2017) propose the use of slang and emoticon dictionaries combined with SentiWordNet and a frequency-based probability classifier in order to obtain the polarity of tweets. Ghiassi, Skinner and Zimbra (2013) builds a domain-specific lexicon the most untactful terms about a topic and, posteriorly use it in a neural network trained to classify the sentiment of tweets of the same topic. Ortigosa, Martín and Carro

(2014) propose a hybrid approach to classify Facebook posts in Spanish. The enriched an existing sentiment lexicon with slangs and recognizing word form variations, like (bueno, buena e buenísimo). They report that an hybrid approach, using SVM algorithm, performs better than a lexicon-based approach.

3.2 Multilingual Sentiment Analysis

Multilingual Sentiment Analysis consists in perform the tasks regarding Sentiment Analysis in datasets containing reviews written in two or more languages. According to Liu (2012), the motivation behind Multilingual Sentiment Analysis is (i) Track the opinion over many countries around the world simultaneously, and (ii) Take advantage of languages with many resources, like English.

The techniques employed for Multilingual Sentiment Analysis fit in the same categories of the Monolingual Sentiment Analysis – Lexicon Based, Machine Learning Based, and Hybrid Techniques. Some researchers deal with multilingualism by employing translation of the data or the resources. Other researchers focus on extracting language independent features from the training data.

Translation-based Approaches rely on machine translation systems in order to translate the documents or some resource (dictionary, lexicons), usually English, to a target language. Then, monolingual techniques of sentiment analysis can be applied to these data. This category of approaches is used to build sentiment analysis systems in languages with few language resources, taking advantages of the richest ones. Can, Ezen-Can and Can (2018) trains a deep learning model based on Recurrent Neural Network to classify polarity of reviews in English. Their training dataset has more than 9 million product reviews to build a classifier, and enrich that model with domain-specific reviews (the authors' focus was on restaurant reviews). That model was used to classify restaurant reviews of four languages – Spanish, Turkish, Dutch and Russian, translated to English by Google Translator API. Araujo et al. (2016) compared translation of the dataset and application of an English Sentiment Analysis technique against language specific approaches. Their study was performed on twitter messages in nine languages (Arabic, Dutch, French, German, Italian, Portuguese, Russian, Spanish, and Turkish). They tested 21 approaches for English Sentiment Analysis and two language-specific techniques. Their results show that in spite of machine translated datasets having produced worse results compared with English datasets, the results achieved are better then the results produced by language-specific

methods.

Another use of Machine Translation techniques can be seen in Mihalcea, Banea and Wiebe (2007). They build a parallel corpus of English and Romanian by translating an English corpus and they projected the annotations of the original corpus into the translated one. Then, a statistical classifier was trained on Romanian corpus. They found that the projection performs better preserving subjectivity of the sentences compared with the translation of the corpus. Mohammad, Salameh and Kiritchenko (2016) investigated whether it is better to translate the documents or the resources in an Arabic-English Sentiment Analysis system. Their experiments involved translate the Arabic Documents to English, manually and using a machine translation software and translate English sentiment lexicons to Arabic. Their results found that automatically translated texts of Arabic running in an English Sentiment Analysis system achieve better results.

Balahur et al. (2014) pointed out that the combination of multilingual data can improve the results of a sentiment classifier. The original training dataset written in English was translated to multiple languages. After that, a sentiment classification technique was applied in each monolingual dataset, in pairs of bilingual datasets, and in a multilingual dataset with all languages togheter. The results shown that the sentiment classification process can be improved by using translated data. Becker, Moreira and Santos (2017) extends the former study, in order to classify emotions instead of polarity on news headlines. Three experiments was designed to test the performance of multilingual data in that task. The use of monolingual datasets, the use of multilingual datasets, and a stack of monolingual classifiers used to build a meta-learner. They conclude that the stack of monolingual classifiers presents the best results, but first it is necessary to find the best configuration of the parameters from the monolingual classifiers composing the meta-learner.

Some researchers, trying to avoid translation, explored features of the multilingual documents in their machine learning models. Nguyen and Nguyen (2018), for example, trained a Convolutional Bidirectional Long Short-Term Memory model. They use word embeddings and contextual information to determine the polarity of Youtube comments. They also use the model to identify the target of the opinions, the product or the video itself. Tellez et al. (2017) proposed an extensive list of features that can be used to polarity assignment. These features were grouped in two sets – the cross-lingual and the language dependent. The cross-lingual features include removal of repeated symbols (for example, *"I loooove this place!!!"*, turns into *"I love this place!"*), removal of diacritic symbols, like accentuation, emoticon handling, grouping then by expressed emotion, control of num-

bers, URLs, users and letter normalization in lowercase. The language dependent features used are stemming, stopwords removal, negation handling (which consists in negate the word that is nearest to the negation, avoiding pronouns and prepositions. For example, in the sentence *"This food was not as good as I expected"* the word *not* will be attached to the word *good*). They also proposed the use of two kinds of tokenizers – n-words (the sentences are tokenized in unigrams and bigrams) and q-grams (which consists in dividing the text in sequences of q characters *e.g.,* the expression *delicious food* divided in 3-grams will produce the tokens {del, eli, lic, ici, cio, ous, us_, s_f, _fo, foo, ood}, where the caracter "_" represents the white space). They use an SVM classifier and two hyper-parameter optimization algorithms in order to select the best subset of features that produce the best results. They achieve good results in their tests in five datasets (Arabic, German, Portuguese, Russian, and Swedish).

3.3 Aspect-Based Sentiment Analysis

The aspect level appears as the most important in Sentiment Analysis, mainly due to the relevant information that it conveys (LIU, 2012). In this level of analysis, the aspects and entities are identified in natural language texts. The aspect phrase extraction task can be classified into three main groups according to the underlying approach (ZHANG; LIU, 2014): (i) based on language rules (HU; LIU, 2004; QIU et al., 2011; PORIA et al., 2014), (ii) based on sequence labelling models (JIN; HO; SRIHARI, 2009; JAKOB; GUREVYCH, 2010), and (iii) based on topic models (MOGHADDAM; ESTER, 2011). However, other works do not fit in only one of these groups as they combine resources from more than one approach (TOH; SU, 2015; TOH; SU, 2016). Furthermore, state-of-the-art approaches rely on more sophisticated architectures like recurrent neural networks such as LSTM, Bi-LSTM, Neural Attention Models, and Convolutional Neural Networks (WANG et al., 2016; WANG et al., 2017; GIANNAKOPOULOS et al., 2017; HE et al., 2017; PORIA; CAMBRIA; GELBUKH, 2016).

Most of the existing work on aspect extraction is designed to deal with reviews written in English. However, in the last few years, researchers started to explore aspect extraction in other languages. In 2016, SemEval made available multilingual datasets for sentiment analysis at aspect level. Participants could use one or more languages in their solutions. The evaluation campaign received 245 submissions from 29 teams for that task (PONTIKI et al., 2016). Such datasets boosted the research in this area. For

example, García-Pablos, Cuadros and Rigau (2018) proposed a topic modeling solution for multilingual aspect extraction and classification, which is almost unsupervised (requiring only a few seed words per language) and achieves competitive results. Most of the works in that task rely on deep learning models. No work achieve best results in all languages. Most of works were tested only in one or two datasets. None of the teams tries to use the datasets in a multilingual way, training a classifier with all datasets together.

The results of the ABSA algorithms are commonly used to produce summaries to show the general opinions contained in a set of reviews. This field of study are know as Aspect-based Opinion Summarization. Liu (2012) mentions that this kind of representation can easily express the sentiment over each one of the aspects of a given target, and also has a quantitative side, where one can perceive the number of users opining on a certain aspect.

Condori and Pardo (2017) proposed two different Aspect-based Opinion Summarization approaches. The first one is an extractive approach, which uses text segments from the review set in order to produce a summary. The second one is an abstractive technique that produces new pieces of text which summarizes the reviews. The extractive algorithm groups the sentences according their aspects and polarity. Next, the algorithm ranks the sentences of each cluster in order to display to the user the ones with best ranking value. The abstactive method is based on pre-built templates, which can generate a summary based on a set of aspects with their opinions. Their approach adopt a different plan according the distributions of the opinions, it means they detect when most of users like or dislike some aspects, or when the users have controversial opinions about a feature, producing a different summary for each situation.

Aspect-based Opinion Summarization approaches can be improved if the set of aspects resulting from an ABSA algorithm were grouped according to their semantic meaning. Next section presents an study of works in aspect clustering field.

3.4 Aspect Clustering of Monolingual Aspect Phrases

After performing the aspect extraction task, clustering aspects is necessary to group together the different representations of the same aspect (*e.g., price*, *cost*, and *charged amount* all refer to the same aspect of a given product). Next, we report on existing approaches for aspect clustering.

Dictionaries and Taxonomies: The first approaches for aspect clustering relied

on pre-existing resources such as Synonym Dictionaries (LIU; HU; CHENG, 2005) or Taxonomies (CARENINI; NG; ZWART, 2005). Dictionaries are usually not considered as good resources for this task because they can not map contextual similarity between expressions. Also, many aspects such as brand names, places or domain-specific words do not typically appear in dictionaries. Taxonomies have the disadvantage of being domain dependent and are difficult to build and maintain. Therefore, these approaches are no longer used.

Topic-Modeling Techniques: The algorithms in this category employ techniques such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), in order to group similar aspect phrases taking into account the semantic similarity between aspects. Guo et al. (2009) proposed a multilevel LSA (mLSA) approach which builds two LDA models in order to group product aspect phrases. It requires semi-structured reviews (with pros and cons) to work. Zhai et al. (2011b) proposed a modification on the original LDA method to support soft constraints like must-links and cannot-links.

Subsequent work has shown that this set of techniques performs poorly. This occurs because this type of technique depends only on probabilistic models based on the frequency of word co-occurrence, which is insufficient to identify the semantic similarity between the aspect phrases.

Semi-Supervisioned Algorithms: This technique was widely used for aspect clustering. It consists of labeling part of the input data with the cluster information, to facilitate and improve cluster formation. In most cases, this information is obtained automatically from the data.

The seminal work in this category is by Zhai et al. (2011a) who automatically obtain the labeled data by leveraging lexical similarity and contextual information. The aspect phrases are first grouped according to the words they share. For example, *"cake"*, *"chocolate cake"*, and *"lava cake"* are joined together in the same cluster in this phase. Next, the lexical similarity between the groups formed in the previous phase is measured in order to group the n most similar groups (n is a hyperparameter of the clustering algorithm). Also, the contextual information about the aspects is taken into account through the use of virtual documents, which consider a window of $[-t, t]$ words before and after every aspect phrase occurrence. Once part of the data is labeled, the EM algorithm based on a Naïve Bayes classifier is employed to cluster the aspect phrases.

Subsequent works changed some characteristics of Zhai et al. (2011a) proposal. For example, there are variations of the pre-grouping heuristics, such as taxonomies auto-

matically extracted from e-commerce pages (WANG et al., 2013), statistical distribution of data in pros and cons reviews (ZHAO et al., 2014), and co-occurrence of aspects and words in reviews (ZHANG; LIU; XIA, 2015). Different algorithms were also used in this task, like k-means (LIU; LV; WANG, 2012; XIONG; JI, 2016), Multinomial Naïve Bayes (ZHAO et al., 2014), and Spectral Clustering (HUANG; NIU; SHI, 2013).

Xiong and Ji (2016), presents a weighted version of virtual documents, built with a semantic relevance metric based on word embeddings. They also propose that the initial constraints of semi-supervised algorithms are strong and can not be violated. Therefore, they measure the degree of belief between the aspect phrases that share some word. That measure is considered in their clustering algorithm and can be domain sensitive. The must-links are used in their flexible-constraint k-means, which is a modification of the original algorithm. Their results show an improvement of the results of Zhai et al. (2011a).

Other Clustering Aproaches: There are also a few proposed solutions to address the problem of clustering aspect phrases which do not fit into the categories discussed in the previous sections. For example, some works (PAVLOPOULOS; ANDROUTSOPOULOS, 2014; ZHAO; QIN; LIU, 2014; HE et al., 2015) use hierarchical clustering in order to produce multi-granular summaries, which can be customized according to the user's needs. Cao, Huang and Zhu (2015) clustered aspect phrases and opinion words simultaneously by using a constrained hidden Markov random field model. Jiajia et al. (2016) combined a feature-opinion relation matrix with two constraint matrices in their clustering model. Finally, Vargas and Pardo (2018) rely on linguistic resources, in order to extract relations between aspect phrases, such as synonym, hypernym, meronym, coreference resolution, causative, deverbal, diminutive/augmentative, foreignism and substring relations.

All algorithms mentioned in this section focused on grouping together monolingual aspect phrases. They cannot be applied to the task of multilingual aspect clustering, because most of the techniques rely on the co-occurrence of context words, but when the reviews are in many languages, the intersection between vocabularies is (almost) empty. We also point out that using dictionaries (or translation) does not perform well in our task. For that reason, we investigate the existing techniques that aim to cluster multilingual documents.

3.5 Multilingual Document Clustering

The task of Multilingual Document Clustering aims to group documents written in more than one language according to their subjects. It is done following two steps. First, the documents in the collection are represented in a language-independent way, and then the groups are formed based on document representations (MA; ZHANG; HE, 2016). The works in this area were developed mostly for grouping news articles, which are typically longer documents. So far, the application of these techniques for clustering reviews (or aspects extracted from reviews) remains unexplored.

This task differs from text classification. In text classification, there are some predefined categories in which the documents are labeled. In document clustering, we do not know the class of the documents *a priori*. For some languages, which are poor in linguistic resources for text processing, there is no labeled data available to perform text classification. In addition, to achieve good results, large amounts of labeled data are required. Labeling data is time-consuming and represents a serious bottleneck.

Multilingual Document Clustering can be considered a harder task compared with Monolingual Document Clustering because the documents in different languages do not share the same vocabulary. Most techniques developed to solve monolingual document clustering rely on the co-occurrence of words in documents. For that reason, most of the state-of-the-art in monolingual document clustering algorithms do not fit into the multilingual configuration of the problem.

We can separate the approaches used to perform multilingual document clustering into two groups, based on the form of representation of the documents in a feature space – monolingual or multilingual feature space.

3.5.1 Monolingual Feature Space Techniques

The approaches in this category aim to create a monolingual feature space of documents in order to cluster multilingual documents. One can use machine translation techniques to translate entire documents or just some document features, while others can rely on multilingual resources like dictionaries or ontologies to create the monolingual feature space. Once this feature space is created, monolingual document clustering techniques can be applied in order to obtain the groups of documents.

A simpler solution for multilingual clustering is to employ machine translation sys-

tems in order to translate the documents, obtaining a monolingual feature space. Flaounas et al. (2011), for example, translate European news in 21 languages to English in order to cluster them together in specific topics.

Other approaches, instead of translating the entire document, select certain features to translate, reducing the effort required to create a monolingual environment. Rauber, Dittenbach and Merkl (2001) eliminate stopwords and infrequent words in the documents before translation. Chen and Lin (2000) translate verbs, nouns and named entities from Chinese to English in order to measure the similarity of documents in these languages.

However, translating the documents in a collection tends to be a costly task. According to Leek et al. (2000), a multilingual document cluster translation-based system leads to around 50% performance loss compared to the monolingual version of that same system. Duek and Markovitch (2018) points out that despite the recent advancement on machine translation systems with the use of deep learning techniques, this kind of software does not have the deep semantic understanding of the documents needed to achieve high-quality translation.

Besides machine translation systems, other multilingual resources can be used in order to generate a monolingual space of documents. For example, one can employ multilingual dictionaries in order to map words between languages. Mathieu, Besançon and Fluhr (2004) represent documents in an adapted version of the vector space model to create monolingual spaces and use bilingual dictionaries in order to map these language spaces. They use a cosine-like similarity measure that takes into account translated words in the tf-idf calculation. Their results reach good purity levels, but low recall. In Hong et al. (2017), bilingual dictionaries are used in order to measure the semantic correlation between Chinese and Vietnamese news, based on the co-occurrence of news elements (entities, verbs, and nouns). It achieved good results in event-centered news clustering.

The use of dictionaries for document clustering has some drawbacks. The polysemy problem is common since words in dictionaries tend to have more than one translation. Thus, a strategy to solve this problem is necessary. Another problem with dictionaries is that they are not available for all language pairs.

A multilingual thesaurus can be used to address the task of creating a monolingual feature space for multilingual clustering. Pouliquen et al. (2004) allied the use of a thesaurus called Eurovoc jointly with independent language features (cognates and geographical references) in order to group together news crawled from the Web. But, this kind of resource is hard to construct and even more scarce then dictionaries.

3.5.2 Multilingual Feature Space Techniques

This category of approaches attempts to map all the documents in a shared language-independent space or to extract language-independent features from the multilingual documents. They also can combine both strategies to represent the document collection (MONTALVO et al., 2006b).

Named Entities are important features used to identify groups in multilingual documents. Montalvo et al. (2006a) measure the named entities similarity using Levenshtein edit-distance function. It allows finding syntactically similar named entities in the documents, for example, their technique identifies *"Barack Obama"*, *"President Obama"* and *"Mr. Obama"* as the same named entity. They employ a heuristic based on the number of related named entities shared by two documents. In Montalvo et al. (2006b) two new clustering algorithms were proposed. The first creates monolingual clusters and then merges them to obtain multilingual clusters. The second algorithm clusters documents directly, which proved to be a better approach. Compared to Montalvo et al. (2006a), the results are slightly worse.

The main disadvantage of these techniques, based on named entity extraction, is that they work better in related languages (like romance languages) and it does not work when the languages have completely different alphabets, like English and Greek or Chinese. Montalvo et al. (2006a) and Montalvo et al. (2006b) perform tests using a collection with English and Spanish documents. There were no tests with languages with more unrelated vocabularies.

A comparable corpus was used by Yogatama and Tanaka-Ishii (2009) to create must-link constraints between monolingual spaces of documents and a multilingual corpus. A propagation algorithm was proposed in order to merge the documents monolingual spaces by propagating the similarity measure between two documents to its nearest neighborhoods.

Denicia-Carral et al. (2010) propose a syntactic similarity measure between words in different languages. For example, the words *president* (in English) and *presidente* (in Portuguese) will be very similar according this measure. After extracting similar word pairs from the documents, they extract thematically related words based on co-occurrence from the context of syntactically similar pairs. For example, from the pair *president* (EN) - *presidente* (PT) we can derive *president-eleições*, *president-candidato*, *presidente-voters*, *etc.*. This technique has the same drawback as the ones based on named entities – the

languages need to have many related words in order to achieve good results.

Topic Modeling Techniques were also employed in the multilingual document clustering task. Wei, Yang and Lin (2008) creates a Language Independent Latent Semantic Indexing Space from a parallel corpus of abstracts of dissertations and theses in English and Chinese. From this, they used monolingual clustering algorithms to categorize another multilingual collection of the same document type.

The focus of multilingual document clustering has been to group news articles. However, review texts differ from news in many ways, especially regarding the size of the documents and the features considered in clustering process. To identify groups of news the most important features tends to be nouns, noun phrases and named entities. We rely on contextual information in order to represent our aspect phrases. We chose to not to translate the multilingual documents, but create a language-independent representation of them instead. Techniques for multilingual document clustering that extract multilingual features usually rely on dictionaries or comparable corpora in order to represent documents. This kind of resource are not available for most languages-pairs and the existing resources are out of domain, limiting its applicability.

3.6 Summary

In this chapter, we analyzed works related to this thesis. Those are divided into two topics – Sentiment Analysis and Multilingual Document Clustering. Based on this knowledge, the next chapter will present our Multilingual Aspect Clustering technique. To the best of our knowledge, this is the first work to address multilingualism in the Aspect Clustering task.

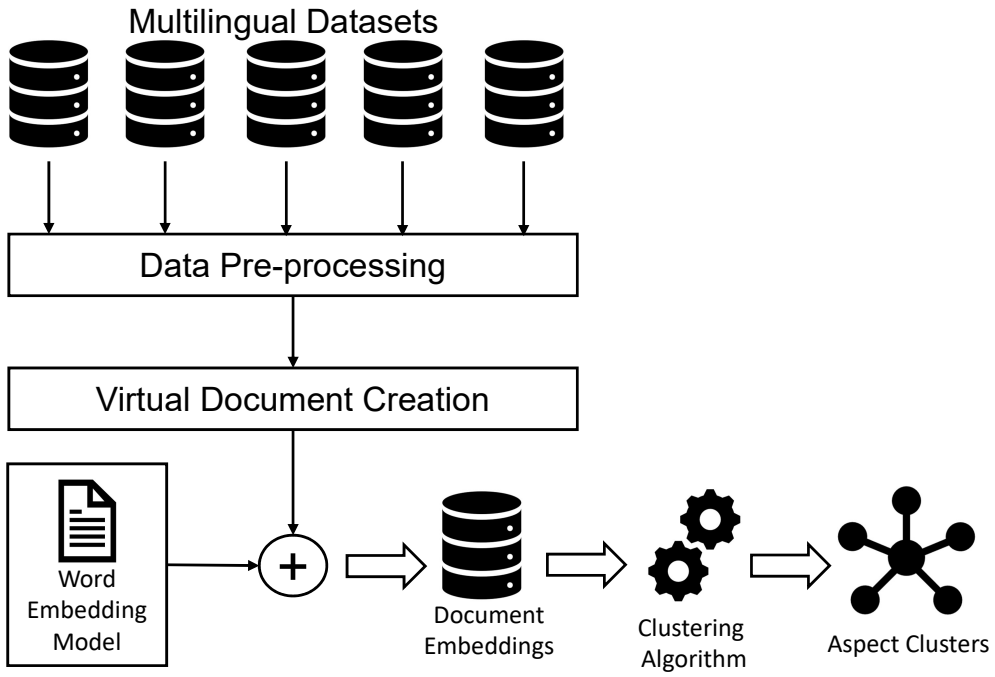
4 MULTILINGUAL ASPECT CLUSTERING

This chapter presents our approach to addressing the problem of Multilingual Aspect Clustering. We start by giving an overview of the solution, and then, each step of the Multilingual Aspect Clustering technique is described in detail.

4.1 Overview

In this work, we apply techniques inspired by multilingual document clustering to (monolingual) aspect clustering in order to address multilingual aspect clustering. Our proposed approach leverages the contextual information of aspect phrases and word embeddings in an unsupervised clustering algorithm in order to group multilingual aspect phrases. Figure 4.1 shows our proposed framework. An important difference between our work and existing proposals, especially in relation to Zhai *et al.* (ZHAI et al., 2011a), is the use of unsupervised learning, which does not require a labeling step before clustering aspect phrases. This makes our method completely automatic and domain independent.

Figure 4.1: The proposed approach for Multilingual Aspect Clustering



Source: The author

4.2 Pre-processing

The input for our method is a set of multilingual reviews along with the extracted aspect phrases. In this work, we assume that the aspects have already been extracted beforehand, so we have not focused on the aspect extraction task. Aspect phrase extraction is outside the scope of this work and can be achieved by the techniques presented in Section 3.3.

We start with a preprocessing phase that consists of three standard steps: (i) splitting of the review text into sentences; (ii) tokenization; and (iii) converting all words to lowercase.

4.3 Virtual Document Creation and Document Embeddings

Once the data is preprocessed, the virtual documents for each aspect phrase in the datasets are built. This step follows the proposal by Zhai *et al.* (ZHAI et al., 2011a), and consists in extracting the contextual information for each aspect phrase present in the review set. The context consists of the surrounding words in a $[-t, t]$ window, removing stopwords and other aspect phrases that co-occur in the same sentence. The Virtual Document of an aspect phrase is the concatenation of the surrounding words of all occurrences of that aspect phrase in the dataset.

For example, in the sentence *"The service is amazing and the ambience is good for a date"*, with a window size of $t = 5$, the Virtual Document of the aspect phrase *service* is composed of the word *{amazing}* and for the aspect phrase *ambience* the Virtual Document has *{amazing, good, date}*. Note that we remove the stopwords *the*, *is*, *and*, *for*, and *a* after the Virtual Documents are constructed. Any aspect phrases that co-occur in the same sentence are also removed from the Virtual Document after its construction *i.e., service* will not be in the Virtual Document for *ambience* and vice-versa.

At the end of the Virtual Document Creation phase, we have l sets of aspect phrases and their respective documents, where l is the total number of languages present on the datasets. In order to group the aspect phrases together, we need to generate a common representation for the aspect phrases that should be language independent. Therefore, we employ *multilingual word embeddings* to this task. Since our documents will be formed by reviews in different languages, it is necessary that the word embeddings can handle this kind of data. For that, one can use embeddings trained with multilingual data, with par-

allel or comparable corpora, or employ techniques that can transform monolingual word embedding spaces into compatible multilingual ones (RUDER, 2017).

The *Document Embeddings* will be a set formed by the union of the word embedding representations of each word in the Virtual Document and the word embeddings of each word in the aspect phrase. In the example above, the Document Embedding of the aspect phrase *service* will be the word embeddings representation of the words *amazing* and *service*. For the aspect phrase *ambience*, its Document Embedding will be the word embeddings of the terms *{amazing, good, date, ambience}*.

4.4 Clustering Document Embeddings

The last step in our approach is to cluster the Document Embeddings in order to group together aspect phrases with the same semantic context. We employ a centroid-based clustering algorithm to that task. This task can be categorized as unsupervised learning because we only use the Document Embeddings as input for our clustering algorithm. We do not use semi-supervised algorithms in our approach, due to the fact that we work with multilingual data. Reviews in multiple languages do not fit into the existing labeling techniques proposed for semi-supervised approaches for monolingual aspect clustering, which focus on the lexical similarity between aspects. Another important fact that makes us use an unsupervised approach is that we do not want to rely on translation of the reviews, neither on the manual labeling of the aspect phrases (which is a difficult task to perform with multilingual data). The goal is to make our approach as simple as possible. Unsupervised approaches are a good choice because they allow us to build a domain-independent solution for multilingual aspect clustering problem.

The pseudocode of our approach is shown in Algorithm 3. The inputs for the algorithm are the Document Embeddings and the desired number of clusters (k). The output is a set of clusters of related Document Embeddings. It initially selects k documents as the centroids (lines 3-5). This selection can be made randomly, or by using some heuristic. We perform experiments selecting the centroids randomly and choosing the aspect phrases that appear more frequently in the dataset as centroids.

Then, for the remaining documents, their distance is measured in relation to the centroids, and the document is assigned to the nearest cluster (lines 7-12). When all documents have been assigned to a cluster, new centroids are chosen, and the documents are re-assigned (lines 13-15). This process converges when, for an entire iteration, no docu-

ALGORITHM 3: Multilingual Aspect Clustering (MAC)

Input : k – Number of Clusters
DE = $DE = \{de_1, de_2, ..., de_i\}$ – Document Embeddings
Output : $DC = \{c_{de_1}, c_{de_2}, ..., c_{de_i}\}$ – Set of Document Clusters

$DC \leftarrow \emptyset$
centroids $\leftarrow \emptyset$
for $i \leftarrow 0$ **to** k **do**
  $centroids \leftarrow$ cent_selection(DE)
end
while *not convergence* **do**
  for $i \leftarrow 0$ **to** $|DE|$ **do**
    for $j \leftarrow 0$ **to** $|centroids|$ **do**
      $distance \leftarrow$ WMD(DE[i], centroids[j])
    end
    $c_{de}[i] \leftarrow$ cluster number with the lowest distance
  end
  for $j \leftarrow 0$ **to** $|centroids|$ **do**
    centroids[i] $\leftarrow$ Result of Equation 4.1
  end
end
return DC

ment changes cluster.

The number of expected clusters (k) depends essentially on the characteristics of the dataset and the goals of the analysis. By increasing k, we obtain a finer granularity which may be desirable in some settings. In Section 5.2, we assess how different values of k impact our clustering quality metrics.

The distance measure we use in order to compare two Document Embeddings is the Word Mover's Distance (WMD) (KUSNER et al., 2015). We chose this measure because it can capture the semantic dissimilarity of two context words from two different aspects and it also works with word embeddings.

We designed a centroid selection method for our problem (instead of using the mean of Document Embeddings, for example) because the WMD distance requires two documents for its calculation, which forced us to always have a document embedding as a centroid of our cluster. The new centroid will be the one that has the lowest WMD average in relation to the other Document Embeddings belonging to that cluster. Equation 4.1 shows how the centroid is chosen.

$$centroid(c) = \min_i \sum_{i=0}^{|c|} \frac{1}{|c|} \sum_{j=0}^{|c|} \mathit{WMD}(\mathit{DE}_i, \mathit{DE}_j) \tag{4.1}$$

where $|c|$ corresponds to the number of elements in a cluster.

To illustrate the behavior of our technique, we provide an example of its operation. Consider the dataset presented in Figure 4.2, composed of the union of restaurant reviews in three languages – English, Spanish, and Dutch. We have previously underlined the aspect phrases in each review. The aspect phrases in this dataset are divided into two aspect groups, food (F) and service (S). The gold partition of each aspect is also highlighted in the dataset.

Figure 4.2: Example Dataset

ENGLISH REVIEWS (R_{EN})
r_1 = I have eaten at Saul, many times, the food (F) is always consistently, outrageously good.
r_2 = The food (F) was well prepared and the service (S) impeccable.
r_3 = The service (S) varies from day to day- sometimes they're very nice, and sometimes not.
r_4 = The pizza (F) is overpriced and soggy.
SPANISH REVIEWS (R_{ES})
r_1 = La atencion(S) es muy buena, los camareros(S) estan muy pendientes de uno todo el tiempo.
r_2 = El servicio (S) es muy bueno y la calidad de la comida (F) al mismo nivel.
r_3 = La calidad de las carnes (F) es insuperable y quiero destacar el excelente servicio (S) recibido.
r_4 = Comida (F) excelente asi como su el servicio de camarero (S).
DUTCH REVIEWS (R_{DE})
r_1 = Het personeel (S) is zeer vriendelijk en correct zonder stijf te zijn.
(The staff is very friendly and correct without being stiff)
r_2 = Proberen de mensen (S) jou met de glimlach te helpen waar ze kunnen.
(People try to help you wherever they can with a smile)
r_3 = Zeer vriendelijke bediening (S). (Very friendly service)
r_4 = Het eten (F) was zeer goed verzorgd en een fatsoenlijke portie!
(The food was very well presented in a decent portion)
r_5 = Slechte en onvriendelijke bediening (S), eten (F) was ronduit slecht.
(Bad and unfriendly service, food was completely bad)

Source: The author

The first step is to preprocess the data and create the Virtual Documents. In this example, we will use a context window of $t = 3$ words. For each aspect phrase, we concatenate the surrounding words of each occurrence of that aspect phrase (excluding the stopwords and other aspect phrases). The Virtual Documents of each aspect phrase in the dataset can be seen in Figure 4.3.

The choice of the hyper-parameter t has an important impact on the performance of the algorithm. Small values of t may not allow sufficient contextual information to be

Figure 4.3: Virtual Documents of aspect phrases

food = {times, consistently, well, prepared}
service = {prepared, impeccable, varies, day}
pizza = {overpriced}
atencion = {buena}
camareros = {buena, pendientes}
servicio = {bueno, destacar, excelente, recibido}
comida = {calidad, mismo, nivel, excelente}
carnes = {calidad, insuperable}
camarero = {servicio}
personeel = {zeer, vriendelijk}
mensen = {proberen}
bediening = {vriendelijke, zeer, onvriendelijke, slechte, ronduit}
eten = {zeer, goed, onvriendelijke, ronduit, slecht}

Source: The author

obtained. See the sentence *"The service varies from day to day– sometimes they're very nice, and sometimes not"*, where the virtual document of the aspect service will not contain the opinion words very and nice. On the other hand, a large t value may cause many out-of-context words to be added to the Virtual Document, damaging the clustering process. For example, in the sentence *"The food was well prepared and the service impeccable"*, the Virtual Document of the aspect service will contain the expression prepared, which belongs to the context of the aspect food.

The next step in the clustering process is to create the Document Embeddings. In this example, three sets of word embeddings are necessary, one for each language of the reviews in the dataset. Recall that the three sets of word embeddings need to be normalized in the same vector space. The Document Embeddings of an aspect will be the union of the word embeddings of the aspect phrase and the embeddings of each word in the Virtual Documents. For example, for the aspect phrase *food*, its Virtual Document will be {food, times, consistently, well, prepared}.

When all the Virtual Documents are built, we can apply the clustering algorithm. The first step in k-means is to define the value of k. In this example, we adopt $k = 2$, as it is the number of gold partitions in the dataset. The centroids can be chosen randomly, or according to some heuristic to allow for a better choice. In this example, we will select the centroids according to the following heuristic "Select two aspect phrases between the ones that appear more frequently in the dataset. If more than two aspect phrases have the maximum frequency, then randomly select two of these aspect phrases". The can-

didate set for our example will contain aspect phrases with frequency two in the dataset $\{\text{food}, \text{service}, \text{servicio}, \text{comida}, \text{bediening}, \text{eten}\}$. As we have several ones with the same frequency, we choose randomly the aspect phrases *food* and *servicio* to be the centroids.

$$centroids \leftarrow \{\text{'food'}, \text{'servicio'}\}$$

Once we selected the centroids, we can compare them with the remaining aspect phrases. Table 4.1 shows the WMD measure between the remaining aspects against the two centroids. Because WMD is a dissimilarity measure, the smallest values of WMD indicate the cluster where that aspect phrase will be allocated. Those values are highlighted in Table 4.1. At the end of this step, we can divide the aspect phrases into two clusters, as shown below.

$c_{\text{food}} = \{$ service, pizza, carnes, mensen, eten $\}$

$c_{\text{servicio}} = \{$atencion, camareros, comida, camarero, personeel, bediening $\}$

Table 4.1: WMD distance between aspect phrases and the centroids (1st Iteration)

	food	servicio
service	**0.993**	1.115
pizza	**1.179**	1.259
atencion	1.219	**1.117**
camareros	1.188	**1.121**
comida	1.040	**0.969**
carnes	**1.125**	1.166
camarero	1.265	**1.017**
personeel	1.129	**1.091**
mensen	**1.186**	1.232
bediening	1.162	**1.124**
eten	**1.042**	1.123

At this point, we can redefine the centroids of each cluster, using Equation 4.1. We then select as the new centroid the aspect phrase which has the smallest average distance in the cluster. Tables 4.2 and 4.3 illustrate how we obtain these values. For example, for the aspect *food* in the first cluster, we calculate the average of the WMD measure between it and the other aspects in the same cluster – food, service, pizza, carnes, mensen, and eten. It results in a score of 0.921. We repeat this calculation for each aspect phrase in the cluster. Looking at the last row in Tables 4.2 and 4.3, we see that the aspect phrase *food* remains as centroid of the first cluster and the aspect *camareros* become the centroid of the second one.

The convergence of our algorithm can be evaluated in function of the changes of

Table 4.2: Centroid Redefinition for cluster food (1st Iteration)

	food	service	pizza	carnes	mensen	eten
food	0.000	0.993	1.179	1.125	1.186	1.042
service	0.993	0.000	1.263	1.218	1.227	1.206
pizza	1.179	1.263	0.000	1.155	1.202	1.667
carnes	1.125	1.218	1.155	0.000	1.241	1.127
mensen	1.186	1.227	1.202	1.241	0.000	1.132
eten	1.042	1.206	1.667	1.127	1.132	0.000
AVG(WMD)	**0.921**	0.985	0.994	0.978	0.998	0.946

Table 4.3: Centroid Redefinition for cluster servicio (1st Iteration)

	servicio	atencion	camareros	comida	camarero	personeel	bediening
servicio	0.000	1.117	1.121	0.969	1.017	1.091	1.124
atencion	1.117	0.000	0.841	1.106	1.260	1.144	1.135
camareros	1.121	0.841	0.000	1.074	1.087	1.113	1.140
comida	0.969	1.106	1.074	0.000	1.220	1.140	1.157
camarero	1.017	1.260	1.087	1.220	0.000	1.153	1.220
personeel	1.091	1.144	1.113	1.140	1.153	0.000	0.772
bediening	1.124	1.135	1.140	1.157	1.220	0.772	0.000
AVG(WMD)	0.920	0.943	**0.911**	0.952	0.994	0.916	0.935

centroids: we achieve the convergence when there are no more changes in the clusters. In our example, this condition was not satisfied, as cluster two change its centroid. Therefore, a new iteration of the clustering algorithm should be performed. First, the centroid set must be updated with the clusters obtained previously:

$$centroids \leftarrow \{\text{‘food’}, \text{‘camareros’}\}$$

After that, the distance of the new centroids to the remaining aspects is calculated. The results can be seen in Table 4.4. After that calculation, we obtain a new configuration of the clusters.

$c_{\text{food}} = \{$ service, pizza, comida, carnes, eten $\}$

$c_{\text{camareros}} = \{$atencion, servicio, camarero, personeel, mensen, bediening $\}$

Finally, after reallocating the aspect phrases in the clusters, the centroids can be updated. Tables 4.5 and 4.6 show that process. The new clusters will be the aspect phrases *comida* and *personeel*.

This process will be repeated until the convergence of the algorithm is reached. Subsequent steps of the algorithm's execution will not be presented, as they are basically the repetition of the steps shown so far.

Table 4.4: WMD distance between aspect phrases and the centroids (2nd Iteration)

	food	camareros
service	**0.993**	1.231
pizza	**1.179**	1.182
atencion	1.219	**0.841**
servicio	1.161	**1.121**
comida	**1.040**	1.074
carnes	**1.125**	1.144
camarero	1.265	**1.087**
personeel	1.129	**1.091**
mensen	1.186	**1.112**
bediening	1.162	**1.140**
eten	**1.042**	1.102

Table 4.5: Centroid Redefinition for cluster food (2nd Iteration)

	food	service	pizza	comida	carnes	eten
food	0.000	0.993	1.179	1.040	1.125	1.042
service	0.993	0.000	1.263	1.177	1.218	1.206
pizza	1.179	1.263	0.000	1.180	1.155	1.667
comida	1.040	1.177	1.180	0.000	0.874	1.038
carnes	1.125	1.218	1.155	0.874	0.000	1.127
eten	1.042	1.206	1.667	1.038	1.127	0.000
AVG(WMD)	0.897	0.976	0.991	**0.885**	0.917	0.930

Table 4.6: Centroid Redefinition for cluster camareros (2nd Iteration)

	camareros	atencion	servicio	camarero	personeel	mensen	bediening
camareros	0.000	0.841	1.121	1.087	1.113	1.179	1.140
atencion	0.841	0.000	1.117	1.260	1.144	1.195	1.135
servicio	1.121	1.117	0.000	1.017	1.091	1.232	1.124
camarero	1.087	1.260	1.017	0.000	1.153	1.259	1.220
personeel	1.113	1.144	1.091	1.153	0.000	1.122	0.772
mensen	1.179	1.195	1.232	1.259	1.122	0.000	1.142
bediening	1.140	1.135	1.124	1.220	0.772	1.142	0.000
AVG(WMD)	0.926	0.956	0.958	1.000	**0.913**	1.018	0.933

4.5 Bissecting k-means

In the previous Section, we applied the standard k-means algorithm to cluster multilingual aspects. The limitation is that we have to inform the number of desired clusters as input and, in many of real-life situations, this parameter may not be known. Thus, in this Section, we propose an alternative approach which uses the *Bisecting k-means* (STEINBACH et al., 2000).

With Bisecting k-means, instead of informing the desired number of clusters, one

informs a threshold value (s), which is the maximum number of aspect phrases that will be allowed in each cluster. Algorithm 4 shows the pseudocode of Bisecting Multilingual Aspect Clustering (BMAC).

ALGORITHM 4: Bisecting Multilingual Aspect Clustering (BMAC)

Input : s – threshold that specifies the maximum number of elements in a cluster
$points$ – data points
Output $DC = \{c_{de_1}, c_{de_2}, ..., c_{de_i}\}$ – Set of Document Clusters
:

```
1  DC ← ∅
2  centroids ← ∅
3  data ← DE
4  repeat
5  |  DC ← MAC(k = 2, DE = data)
6  |  c ← cluster with maximum number of data points
7  |  max ← length of cluster c
8  |  data ← data points of cluster c
9  until max ≤ s
10 return DC
```

The BMAC alternative starts applying our MAC algorithm into all data points, with k=2. Once the data is clustered, our algorithm chooses the largest cluster and splits it into two other clusters. We repeat this division (*i.e.,* bisection) process until the size of the clusters is smaller than the threshold parameter.

Considering the example presented in the previous Section, we present a simulation of the execution of BMAC. The parameter s is set to 3. BMAC starts clustering all the 13 data points into two clusters. Assume it resulted in the following clusters:

$\text{DC}_1 = \{$food, service, pizza, comida, carnes, eten $\}$

$\text{DC}_2 = \{$camareros, atencion, servicio, camarero, personeel, mensen, bediening $\}$

In the next iteration, the BMAC algorithm will be applied just in the data points belonging to DC_2, which is the largest cluster, that leads to the emergence of a new cluster. In this point, the aspect phrases in the dataset are grouped into three clusters. We present below a possible configuration of the aspect clusters.

$\text{DC}_1 = \{$food, service, pizza, comida, carnes, eten $\}$

$\text{DC}_2 = \{$camareros, camarero, personeel $\}$

$\text{DC}_3 = \{$atencion, servicio, mensen, bediening $\}$

The algorithm chooses now cluster DC_1 to split. After that, the cluster set will contain four clusters.

$\text{DC}_1 = \{$food, pizza, comida, carnes, eten $\}$

$\text{DC}_2 = \{$camareros, camarero, personeel $\}$

$\text{DC}_3 = \{$atencion, servicio, mensen, bediening $\}$

$DC_4 = \{\text{service}\}$

This process continues until all clusters have at most three aspect phrases. Thus, clusters DC_1 and DC_3 need to be divided at least one more time. It is important to notice that the value of the hyper-parameter s impacts in the number of clusters. The smaller the s value, the more fine-grained the clusters will be. However, if the s is too small, it tends to generate an elevated number of groups. We study the impact of this parameter and our findings in Section 5.2.4.

4.6 Summary

This chapter presented an approach to solving the problem of multilingual aspect clustering. We use contextual information of reviews combined with multilingual word embeddings in order to represent the aspect phrases. An unsupervised clustering algorithm – k-means was used to group together related aspect phrases. We also present two version of our clustering algorithm. In the traditional k-means the user must specify the number of desired clusters. Sometimes this information is unknown. Therefore a bisecting k-means version of our algorithm was designed. In that version, the user specifies the maximum number of aspect phrases that will be allowed and the algorithm runs until that condition is true. In the next Chapter, we show an experimental evaluation of our proposed technique.

5 EXPERIMENTAL EVALUATION

In this chapter, we describe the evaluation of our proposed multilingual aspect clustering technique. Initially, the experimental setup is presented and then the results are discussed.

5.1 Experimental Design

The experimental design includes the description of the dataset, the baseline, the evaluation metrics, and the setup used in our approach.

5.1.1 Datasets

We used the Restaurant datasets from SemEval 2016 - Task 5[1] in order to evaluate our approach. This is a multilingual dataset with reviews in five languages: English, Dutch, Russian, Spanish, and Turkish. It was originally designed for the aspect extraction task. Some statistics of the datasets can be found in Table 5.1.

Table 5.1: Statistics of the SemEval Datasets

Dataset	#Reviews	#Sentences	#Aspect Phrases
English	350	2,000	644
Dutch	300	1,722	508
Russian	312	3,655	1,024
Spanish	627	2,070	543
Turkish	300	1,232	831
Total	1,889	10,679	3,550

Figure 5.1 shows an entry of a review in SemEval Dataset. Reviews are delimited by the tag <Review>. Each review is divided in sentences. Every sentence in this dataset has annotations about the aspect phrases classified into six aspect clusters: Restaurant, Food, Drinks, Service, Ambience, and Location. We used this classification scheme as the gold standard in our evaluations. The dataset also provides information about the polarity and the localization of the aspect phrase in the sentence (parameters from and to). The annotations in this dataset include explicit and implicit aspects. The implicit aspects have the property target marked as `NULL` in the dataset. It is important to notice that we

[1] Avaiable at <http://alt.qcri.org/semeval2016/task5/>

only used in our approach the explicit aspect phrases in the reviews, because we can not extract contextual information of the implicit aspect phrases, since their opinion target are annotated as `NULL` in our datasets.

Figure 5.1: Example of Review on SemEval Dataset

```
<Review rid="es_9reinas_2_AlbertMuntana_2015-03-09">
    <sentences>
        <sentence id="es_9reinas_2_AlbertMuntana_2015-03-09:0">
            <text>La verdad es que todo muy bien; el servicio, la
             ↪ comida y la apariencia, todo correcto.</text>
            <Opinions>
                <Opinion target="NULL" category="RESTAURANT#GENERAL"
                 ↪ polarity="positive" from="0" to="0"/>
                <Opinion target="servicio" category="SERVICE#GENERAL"
                 ↪ polarity="positive" from="35" to="43"/>
                <Opinion target="comida" category="FOOD#QUALITY"
                 ↪ polarity="positive" from="48" to="54"/>
                <Opinion target="apariencia"
                 ↪ category="AMBIENCE#GENERAL" polarity="positive"
                 ↪ from="60" to="70"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Source: <http://alt.qcri.org/semeval2016/task5/>

Some aspect phrases are categorized in more than one of the above aspect clusters, so we chose as the category the one with the most assignments for that aspect phrase. We made this decision because we noticed that just a few aspect phrases in our dataset have occurrences in two or more categories – 129 in 3,550 aspect phrases. We also noticed that in most cases, when an aspect phrase has more then one category, there is a major one with many occurrences, while the other categories have very few (one occurrence in general). In most of cases in which an aspect phrase was annotated in more than one aspect group the aspect groups were RESTAURANT, AMBIENCE, and LOCATION. These three categories are quite similar, and people tend to use the same aspect phrases in order to describe them. For example, the aspect phrase *restaurant* was used to describe the restaurant itself – *Best restaurant in Brooklyn*, the ambience *It's a small cute restaurant*, and the restaurant location – *The restaurant looks out over beautiful green lawns to the Hudson River and the Statue of Liberty*. Although it is used quite differently, if we observe the frequency of use of the *restaurant* aspect in the English dataset we see that of the 43 apparitions of this aspect in 33 it referred to the aspect group RESTAURANT, in nine to the AMBIENCE and in just one to the LOCATION cluster.

Another important feature of this dataset that deserves to be highlighted is the distribution of aspect clusters, shown in Table 5.2. Over half of the aspect clusters in our datasets refer to the food cluster. Also, none of the other aspect clusters has more than 20% of the aspect phrases in the review set R. This is a peculiarity of restaurant reviews, in which the food aspect has a huge importance compared to the other aspects in this domain.

Table 5.2: Distribution of the Aspect Clusters in the Reviews

Aspect Cluster	#Aspect Phrases	%
Restaurant	416	11.72
Food	1,828	51.49
Drinks	256	7.21
Service	497	14.00
Ambience	497	14.00
Location	56	1.58
Total	3,550	100

The SemEval Dataset was not designed to the task of aspect clustering. This dataset was used in research in monolingual aspect extraction in which each language is treated separately. The data contains reviews from multiple restaurants, which groups under the same category very different aspect phrases. For example, in food category, we have aspect phrases related to Italian restaurants (ravioli, pizza), Japanese Restaurants (sushi, temaki), French Restaurants (foie gras, gâteau), among many other kinds of restaurants. This excessive variation of aspects poses a challenge for clustering task because the algorithm had to group all these different aspect phrases under the same aspect group. We performed tests with more than six clusters, in order to obtain more specific clusters. We also used the Bisecting k-means approach aiming to fit the data into a more suitable number of clusters.

5.1.2 Baseline

In order to evaluate our technique, we implemented the algorithm proposed by Zhai et al. (2011a) as baseline. This approach was chosen because it is the most seminal paper in aspect clustering field, it is simple, well detailed, and requires few resources in its implementation (Wordnet e lists of stopwords). In addition, it can be applied to our datasets because it does not require semi-structured data or extra manual annotations in the reviews. Despite having been proposed almost seven years ago, this work is still highly cited – having received over 30 citations in 2018 (according to GoogleScholar).

Due to the fact that this technique was originally designed for monolingual aspect

clustering, we had to make some adaptations for it to work with multilingual data. First, we translated the reviews into English. We also removed the stopwords from the virtual documents in two occasions, before and after translation, because we notice that the English stopword list was more accurate than the lists in other languages. Finally, we considered words with the same translations as if they were the same aspect phrases. For example, the words 'nagerecht', 'десерт', 'postre', and 'tatlı' are all grouped together in the same virtual documents of the aspect phrase 'dessert'. The remaining configurations are the same as in the original article, described in Section 3.1.

5.1.3 Evaluation Metrics

As proposed by Zhai *et al.* (ZHAI et al., 2011a), we measured the performance of our clustering algorithms in terms of Entropy and Purity. In our evaluation, we consider a dataset DS, clustered into k disjoint sets $\{DS_1, DS_2, ..., DS_K\}$ and its respective golden partitions $G = \{g_1, g_2, ..., g_K\}$. The goal of the clustering algorithm is to minimize entropy and maximize purity.

Purity: Purity intends to measure the largest portion of a cluster that contains data from a single golden partition *i.e.,* the highest percentage of correctly clustered points. It can be calculated as in Equation 5.1, where $P_i(g_i)$ is the proportion of g_i data points in D_i. The purity of entire clusters is calculated according to Equation 5.2.

$$purity(DS_i) = \max_j P_i(g_j) \tag{5.1}$$

$$purity_{total} = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} purity(DS_i) \tag{5.2}$$

Entropy: The entropy of a cluster is measured by the proportion of each gold partition present in it. It is calculated as in Equation 5.3. The entropy of a cluster is obtained following Equation 5.4.

$$entropy(DS_i) = -\sum_{j=1}^{k} P_i(g_j) \log_2 P_i(g_j) \tag{5.3}$$

$$entropy_{total} = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} entropy(DS_i) \tag{5.4}$$

5.1.4 Multilingual Aspect Clustering Setup

The configuration setup for our proposal is as follows. The Virtual Documents were created considering a window of $[-10, 10]$ words. FastText[2] was used for word embeddings. Their authors have made available the pre-trained multilingual word vectors for 157 languages trained on Wikipedia. We employed their models in order to treat out of vocabulary words, which enriched our review representations. In order to aligning the FastText vectors, we use the transformation matrices of Smith *et al.* (SMITH et al., 2017)[3]. We chose not to train our own word embedding representation because that would require a huge amount of data to achieve a minimally satisfactory word embedding model. Since we are working with reviews, which tent to be short texts, the task is harder – especially for some languages with few reviews available online.

We used the Gensim[4] package for the word vector representations and for computing WMD score. The centroid-based clustering algorithm used was k-means. This choice was motivated by its efficiency (*i.e.,* it does not require pairwise comparisons among all data items) and the fact that we can choose the number of clusters that should be generated.

We present two versions of our multilingual aspect clustering algorithm, the traditional version of k-means and an alternative using the bisecting k-means algorithm. We refer to these versions as MAC and BMAC, respectively.

We also tested two different centroid selection techniques. In MAC-RAND/BMAC-RAND, we chose the centroids randomly, while in MAC-CENT/BMAC-CENT we selected as centroids the aspect phrases that appear more frequently in the datasets. In MAC approaches, we set the value of k to six, as it was the number of gold partitions on our dataset. For BMAC, we set the number of the hyper-paramenter s as 100, which means that our clusters will have a maximum of a hundred aspect phrases each.

The tests were run in each language separately, and with all languages together. In our experiments with MAC-RAND, we ran the algorithm ten times and calculated the averages for purity and entropy to mitigate the effects of variability. As for the MAC-CENT, the results do not change across different runs because the chosen centroids are always the same.

[2] Avaiable at <https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md>
[3] Avaiable at <https://github.com/Babylonpartners/fastText_multilingual>
[4] Avaiable at <https://radimrehurek.com/gensim/>

5.2 Results

This section contains the results of the tests performed in our Multilingual Aspect Clustering algorithm and also presents an analysis of the performance of our technique.

5.2.1 Overall Results

Tables 5.3 and 5.4 present the results of our methods (MAC and BMAC) and the baseline (L-EM). The results show that our BMAC technique outperforms the entropy values of the baseline (where smaller values mean better results). For purity, MAC-CENT achieves the best results for the English dataset, while BMAC surpasses the baseline in the other datasets. Recall that our method is unsupervised, while our baseline relies on two phases of pre-processing before clustering the aspect phrases. Also, the baseline requires all the reviews to be in the same language.

Table 5.3: Experimental Results – ENTROPY

Method	English	Dutch	Russian	Spanish	Turkish	All
L-EM	1.748	1.753	1.974	1.591	2.076	1.932
MAC-RAND	1.654	1.801	1.814	1.517	2.002	$1.858^{\triangle}$
MAC-CENT	1.624	1.719	1.706	1.540	2.039	$1.841^{\triangle}$
BMAC-RAND	1.587	**1.708**	**1.552**	**1.425**	1.959	$\mathbf{1.534}^{\triangle}$
BMAC-CENT	**1.552**	1.716	1.663	1.444	**1.940**	$1.550^{\triangle}$

Table 5.4: Experimental Results – PURITY

Method	English	Dutch	Russian	Spanish	Turkish	All
L-EM	0.598	0.591	0.503	0.644	0.484	0.549
MAC-RAND	0.605	0.559	0.540	0.644	0.487	$0.543^{\triangledown}$
MAC-CENT	**0.629**	**0.576**	0.571	0.581	0.492	$0.539^{\triangledown}$
BMAC-RAND	0.605	0.564	**0.614**	0.648	**0.498**	$0.599^{\triangledown}$
BMAC-CENT	0.613	0.564	0.581	**0.654**	0.492	$\mathbf{0.605}^{\triangledown}$

To evaluate if the difference between the results of our technique and the baseline are significant, we conduct paired t-tests. The results are presented in Tables 5.3 and 5.4. In the following tables, a $\triangle$ symbol indicates that p-value < 0.05, which means a significant difference. The symbol $\triangledown$ indicates that the p-value > 0.05 which indicates the difference is not significant.

On average, BMAC with centroid selection shows the best results for entropy and purity for our multilingual dataset. Their results proved to be more consistent than those

obtained with the random selection of centroids. Bisecting k-means also is more indicated than traditional k-means, because this technique can find a more suitable number of clusters for the dataset.

5.2.2 Analysis of the Resulting Clusters

Table 5.5 shows some excerpts of the clusters generated by our algorithm. Based on these results, we will discuss some strengths and weaknesses of our approach. Cluster number one, for example, shows that our method is able to group together aspect phrases that are synonyms in the same language (*waitress*, *waitstaff*, *servers*), or across languages (управляющий, manager, propietario). We also noticed that our approach is able to detect similar words with different spellings. This phenomenon is frequent in the Russian language, as can be seen in cluster two, where we see many word groups where this property can be validated (рестораном, ресторане and ресторан, for example). This property can be considered as an improvement over the heuristic used by the baseline, which just groups aspect phrases that share equal words. Our approach can also reproduce the effects of this heuristic, an example of that is cluster three, that groups the aspect phrases with the word *"menu"*, which has the same form in English, Dutch, Spanish, and Turkish. It is interesting to note that our algorithm also includes in this cluster the aspect phrase меню, which is the translation of the word menu to Russian.

MAC is able to group together semantically related aspect phrases. We present two examples of this ability in Table 5.5. Cluster four groups aspect phrases related to seafood dishes (which can be seen as the aspect cluster of this group). At the same time, cluster five has aspect phrases related to artistic presentations. We obtained this result thanks to an adequate representation of the virtual documents, allied to a measure of similarity that is strong enough to capture the contextual similarity between aspect phrases.

Despite the good results, our algorithm has some limitations. Because it is an unsupervised approach, it suffers from the drawbacks of this type of algorithm. Sometimes it is hard to guide the learning process in order to reach our clustering goal, which causes some aspect phrases to be misclassified. Another issue is to do with the integration of the multilingual datasets. We noticed that sometimes the clusters have only aspect phrases in one language. This is caused by the bias introduced in the normalization of the word embeddings phase. This can be seen when we make a comparison between the cosine similarity of a word and its translations into other languages. For example, the distance

Table 5.5: Excerpts of clusters generated by our algorithm

#	Centroid	Aspects
1	waitress	управляющий (manager) – waitstaff – hostess – manager – servers – gentleman – propietario (manager)
2	обслуживание (service)	сервис (service) – обслуживания (service) рестораном (restaurant) – ресторане (restaurant) – ресторан (restaurant) – место (place) – заведение (establishment) – заведению (establishment) атмосфера (atmosphere) – атмосферу (atmosphere) – интерьера (interior) – интерьер (interior) официантов (waiters) – официанты (waiters) – официант (waiter) – официантка (waitress) – персонала (staff) – персонал (staff) кухню (kitchen) – кухней (kitchen) – кухня (kitchen) – Качество кухни (quality of kitchen) – кухни (kitchens) музыка (music) – живая музыка (live music)
3	menu 'parels van india' (menu 'pearls of india')	menu kaart (menu card) – 3 gangen menu (3 course menu) – детское меню (children's menu) – блюд из меню (dishes from the menu) – sake menu – Menu de Primavera (Spring Menu) – menu fiyatları (menu prices)
4	scallops	scampi in de look – рыба в беконе (fish in bacon) – sea urchin – fried shrimp – lobster knuckles – oysters – stir fry blue crab – fried oysters and clams – pulpo con langostinos (octopus with prawns) – vieira con sopa (scallop with soup) – soya soslu somon (salmon with soy sauce)
5	грузинские танцевальный коллектив (Georgian dance group)	голос солистки (voice of the soloist) – концертик (concert) – песни (songs) – программа (program) – belly dancing show – müzik seçimleri (music selections) – las Posesas (theater play)

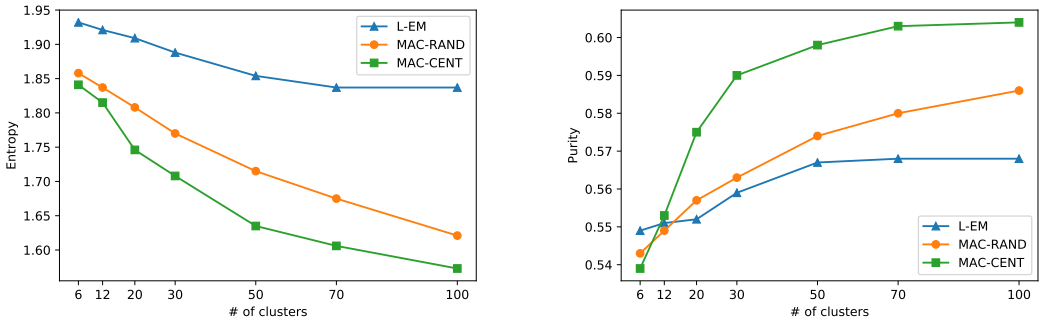
between the vector of the word 'dessert' and its translations десертам, postre, nagerecht and tatlı is 0.68, 0.59, 0.71, 0.48 respectively, while the most similar words in English are desserts (0.91), pastries (0.81), cakes (0.76), pancakes (0.74) and salads (0.74). This happens in cluster two of Table 5.5, which has only aspect phrases in Russian. Some of that aspect phrases are more related to other clusters instead of the cluster two, for instance, музыка (music) and живая музыка (live music) are more related to cluster five.

5.2.3 Variation of parameter k in MAC

To assess how the method behaves with different number of clusters, we performed tests varying the parameter k. The results can be seen in Figure 5.2. When the number of clusters increases, our method showed a more pronounced drop in entropy and a gain in purity, compared to the baseline (Figure 5.2a). This happens because the heuristics of the baseline to label the data tends to get worse as the number of clusters increases. At the same time, our approach tends to select better initial centroids as the number of

cluster increases. For a small number of clusters, our centroid selection technique did not work well, because the more frequent aspect phrases refers to the same aspect cluster. For example, it selects as centroids service, обслуживание музыка and servicio, which are the translation of service for Russian and Spanish. With a more largest number of clusters, the centroid selection algorithm tends to select more diversified aspect phrases.

Figure 5.2: Experimental results with variation of k



(a) Entropy (b) Purity

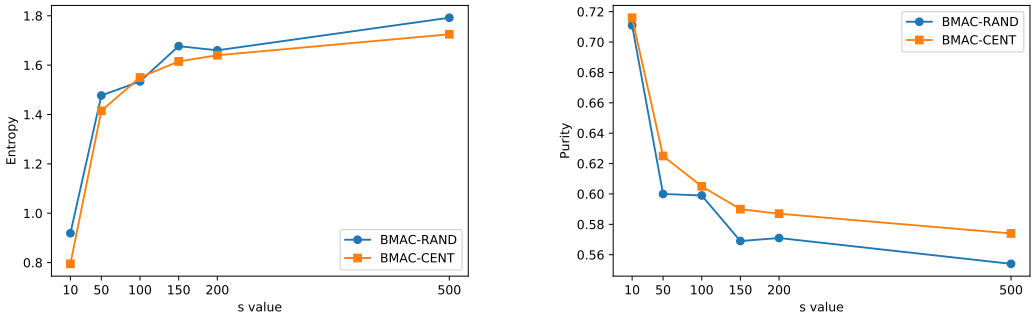
Source: The author

5.2.4 Variation of parameter s in BMAC

In order to understand the behaviour of the paramenter s in BMAC, we perform tests variating its value. BMAC algorithm was executed for the entire dataset, variating the variable s from 10 to 500. The results are shown in Figure 5.3. We notice that the higher the s value, the worse the purity and entropy values tend to be. These measures tend to grow/decrease exponentially as the number of aspects per cluster increase. The figure also shows that random selection of initial centroids tends to turns the results worse, specially when the value of s are bigger. It can be noticed that the centroid selection technique tends to produce more consistent results, considering that their values oscillate less than those produced by the random selection of centroids.

Another variable to consider choosing the value of the parameter s is the number of clusters generated by the BMAC. Depending on their value this technique may create an excessive number of clusters in order to fit in the constraints imposed by the algorithm. Figure 5.4 shows the the number of clusters generated for each execution of BMAC on the 3,550 aspect phrases present in our dataset. The green line in the chart represents the minimum number of clusters that should be generated by the algorithm, it means, the total

Figure 5.3: Experimental results with variation of s



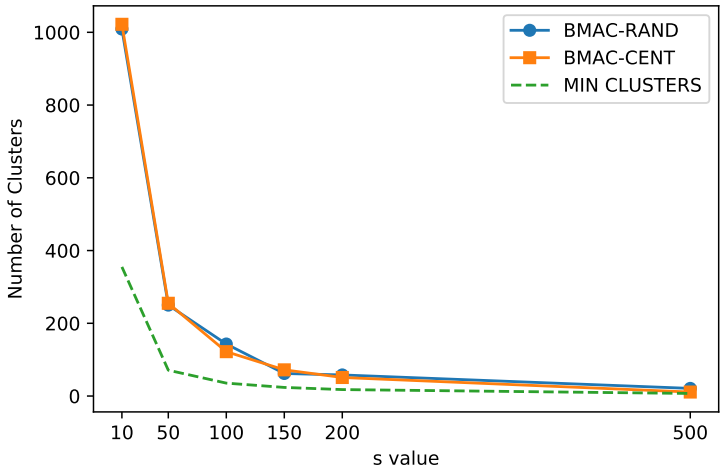
(a) Entropy (b) Purity

Source: The author

of aspect phrases in the dataset divided by the s variable. This variable has to be adjusted based on the size of the dataset.

Larger values of s tend to produce a small number of clusters, but the values of entropy and purity tend to get worse. On the other hand, a small s value produces a large number of clusters, which are not desirable for most of applications. In this work, we chose this variable empirically, by analyzing the results of BMAC produced by the different s values.

Figure 5.4: Number of clusters according the variation the value of s parameter



Source: The author

5.2.5 Time Complexity Analysis

According to Tan et al. (2013) the time complexity of k-means algorithm is given by $O(I * K * m * n)$, where I is the number of iterations to achieve convergence, K is the number of clusters, m corresponds to the number of data points, and n is the number of attributes of each data point. The author also says that the variable m is the most influential in the complexity calculation, concluding that the complexity is linear on the number of data points.

The WMD similarity function used in our algorithm has a polynomial complexity of $O(p^3 * \log p)$ where p is the number of unique words in the documents (KUSNER et al., 2015). This function is used in our centroid selection technique, presented in Equation 4.1. This function has a complexity function of $O(m^2 * K)$. Thus, the variables with most impact in our clustering approach are the number of aspect phrases (m) and the size of the Virtual Documents.

Our MAC and BMAC clustering process take some hours to group the 3,550 aspect phrases in the datasets. The first iterations occupied most of that time, because the changes in clusters tend to decrease over time. We store the results of WMD computation in main memory, which speed up the aspect clustering.

5.3 Summary

This chapter presented the experiments ran in our multilingual aspect clustering technique. We start presenting the design of such experiments, which involves the description of datasets, the baseline implemented, the evaluation metrics, and the setup of our algorithms.

The results shown that our bisecting k-means version of multilingual aspect clustering algorithm allied with the centroid selection heuristic achieve the best results in our multilingual dataset. Analyzing that results we noticed that our technique are good in detecting semantic related aspects. We presented some interesting patterns found in our results. We also pointed some drawbacks of our method, and possible improvements that can be done in the future.

At the end of the chapter, we presented an analysis of how our MAC approach behaves when the parameter k changes. A similar study was done in our BMAC algorithm analyzing the variations on the s parameter.

6 CONCLUSION

In this work, we proposed an unsupervised approach to address the problem of Multilingual Aspect Clustering.

The main contributions of this work include the formulation of the problem of aspect clustering in a multilingual set of reviews, and the presentation of an unsupervised technique to solve that problem, which combines multilingual word embeddings and the WMV similarity measure in order to group together aspect phrases in multiple languages. Because it is unsupervised, this approach can be applied across domains, requiring only word embeddings for each language in the review set. Our technique can be used as baseline for researchers in their works. We carried out experiments on restaurant reviews written in English, Spanish, Russian, Dutch, and Turkish and compared our performance against a established baseline. The results show that our unsupervised clustering technique achieves results that outperforms the results of a semi-supervised baseline. Our experiment were made on a publicly available dataset, while most of the works on monolingual aspect clustering do not make their datasets available .

A paper was written as part of this dissertation. It was published as a regular paper at the International Conference on Web Intelligence 2018 (PESSUTTO; VARGAS; MOREIRA, 2018). This paper contains the description of our multilingual aspect clustering technique and the experiments made with the traditional k-means algorithm. We intend to submit an extended version of this paper with the results achieved by bisecting k-means in a journal.

Future work will include a study to mitigate the weaknesses of our approach (use of a more accurate technique of word embedding normalization and pruning the aspect phrases of the clusters, in order to remove irrelevant ones). Another important feature will be the development of heuristics that can be used with multilingual data and will allow us to use semi-supervised approaches in multilingual aspect clustering. We also want to test our technique in a dataset with reviews of only one restaurant, which would allow us to work with a reduced (and more related) set of aspect phrases, achieve better results, and enable fine-tuning. For that, we need to create and annotate a dataset, because this resource does not currently exist. It is also interesting to test out approach in different domains, like product reviews. Another aspect that can be more deeply explored in the future is the multigranularity of the clusters. It can be done with the use of hierarchical clustering algorithms. However, it will be necessary to make some adaptations in our

technique in order to use that kind of algorithms. For example, in our approach we just compare pairs of document embeddings using WMD measure. Traditional hierarchical clustering algorithms require the comparison between groups of document embeddings.

Aspect-Based Sentiment Analysis is a challenging task, because an opinions can be expressed in many different ways in reviews. There are some specific issues that were not covered by our technique. An example of that is implicit aspects. A modified version of our approach which takes into account implicit aspects would be a desirable improvement. Another issue that deserves a better investigation in future work is aspect ambiguity, which occurs when an aspect phrase are classified in different aspect clusters. This problem was acknowledged in Section 5.1.1. A better understanding of this problem may bring new insights and improve the results achieved by this work. Finally, we wish to build a visualization tool that summarizes the results of aspect clustering and explore the customization in order to emphasize the aspects in which the user has interested.

REFERENCES

ARAUJO, M. et al. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In: ACM. **Proceedings of the 31st Annual ACM Symposium on Applied Computing**. [S.l.], 2016. p. 1140–1145.

ASGHAR, M. Z. et al. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. **PloS one**, Public Library of Science, v. 12, n. 2, p. e0171649, 2017.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval: The Concepts and Technology Behind Search**. [S.l.]: Addison Wesley, 2011. ISBN 9780321416919.

BALAHUR, A.; PEREA-ORTEGA, J. M. Sentiment analysis system adaptation for multilingual processing: The case of tweets. **Information Processing & Management**, v. 51, n. 4, p. 547 – 556, 2015. ISSN 0306-4573. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0306457314000934>.

BALAHUR, A. et al. Resource creation and evaluation for multilingual sentiment analysis in social media texts. In: CITESEER. **LREC**. [S.l.], 2014. p. 4265–4269.

BANEA, C.; MIHALCEA, R.; WIEBE, J. Multilingual subjectivity: are more languages better? In: **International Conference on Computational Linguistics**. [S.l.: s.n.], 2010. p. 28–36.

BECKER, K.; MOREIRA, V. P.; SANTOS, A. G. dos. Multilingual emotion classification using supervised learning: Comparative experiments. **Information Processing & Management**, v. 53, n. 3, p. 684 – 704, 2017. ISSN 0306-4573.

CAN, E. F.; EZEN-CAN, A.; CAN, F. Multilingual sentiment analysis: An rnn-based framework for limited data. **arXiv preprint arXiv:1806.04511**, 2018.

CAO, Y.; HUANG, M.; ZHU, X. Clustering sentiment phrases in product reviews by constrained co-clustering. In: **Natural Language Processing and Chinese Computing**. [S.l.]: Springer, 2015. p. 79–89.

CARENINI, G.; NG, R. T.; ZWART, E. Extracting knowledge from evaluative text. In: ACM. **Proceedings of the 3rd international conference on Knowledge capture**. [S.l.], 2005. p. 11–18.

CHEN, H. et al. Neural sentiment classification with user and product attention. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2016. p. 1650–1659.

CHEN, H.-H.; LIN, C.-J. A multilingual news summarizer. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 18th conference on Computational linguistics-Volume 1**. [S.l.], 2000. p. 159–165.

CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. **Expert Systems with Applications**, v. 78, p. 124 – 134, 2017. ISSN 0957-4174. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0957417417300829>.

DENICIA-CARRAL, C. et al. Bilingual document clustering using translation-independent features. In: **Proceedings of CICLing**. [S.l.: s.n.], 2010. v. 10.

DOU, Z.-Y. Capturing user and product information for document level sentiment analysis with deep memory network. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2017. p. 521–526.

DUEK, S.; MARKOVITCH, S. Automatic generation of language-independent features for cross-lingual classification. **CoRR**, abs/1802.04028, 2018. Available from Internet: <http://arxiv.org/abs/1802.04028>.

ERTUGRUL, A. M.; ONAL, I.; ACARTURK, C. Does the strength of sentiment matter? a regression based approach on turkish social media. In: SPRINGER. **International Conference on Applications of Natural Language to Information Systems**. [S.l.], 2017. p. 149–155.

FELDMAN, R. Techniques and applications for sentiment analysis. **Commun. ACM**, ACM, New York, NY, USA, v. 56, n. 4, p. 82–89, abr. 2013. ISSN 0001-0782. Available from Internet: <http://doi.acm.org/10.1145/2436256.2436274>.

FLAOUNAS, I. et al. Noam: news outlets analysis and monitoring system. In: ACM. **Proceedings of the 2011 ACM SIGMOD international conference on Management of data**. [S.l.], 2011. p. 1275–1278.

GARCÍA-PABLOS, A.; CUADROS, M.; RIGAU, G. W2vlda: almost unsupervised system for aspect based sentiment analysis. **Expert Systems with Applications**, v. 91, p. 127–137, 2018.

GHIASSI, M.; SKINNER, J.; ZIMBRA, D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. **Expert Systems with applications**, Elsevier, v. 40, n. 16, p. 6266–6282, 2013.

GIACHANOU, A.; CRESTANI, F. Like it or not: A survey of twitter sentiment analysis methods. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 49, n. 2, p. 28:1–28:41, jun. 2016. ISSN 0360-0300. Available from Internet: <http://doi.acm.org/10.1145/2938640>.

GIANNAKOPOULOS, A. et al. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. In: **Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 180–188.

GUO, H. et al. Product feature categorization with multilevel latent semantic association. In: ACM. **Proceedings of the 18th ACM conference on Information and knowledge management**. [S.l.], 2009. p. 1087–1096.

HE, R. et al. An unsupervised neural attention model for aspect extraction. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 388–397.

HE, Y. et al. Clustering chinese product features with multilevel similarity. In: **Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data**. [S.l.]: Springer, 2015. p. 347–355.

HONG, X. et al. Cross-lingual event-centered news clustering based on elements semantic correlations of different news. **Multimedia Tools and Applications**, Springer, v. 76, n. 23, p. 25129–25143, 2017.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. New York, NY, USA: ACM, 2004. p. 168–177.

HUANG, S.; NIU, Z.; SHI, Y. Product features categorization using constrained spectral clustering. In: MÉTAIS, E. et al. (Ed.). **Natural Language Processing and Information Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 285–290.

IOSIFIDIS, V.; NTOUTSI, E. Large scale sentiment learning with limited labels. In: ACM. **Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.], 2017. p. 1823–1832.

ISMAIL, H. M.; BELKHOUCHE, B.; ZAKI, N. Semantic twitter sentiment analysis based on a fuzzy thesaurus. **Soft Computing**, v. 22, n. 18, p. 6011–6024, Sep 2018. ISSN 1433-7479. Available from Internet: <https://doi.org/10.1007/s00500-017-2994-8>.

JAKOB, N.; GUREVYCH, I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In: **Proceedings of the 2010 conference on empirical methods in natural language processing**. Massachusetts, USA: Association for Computational Linguistics, 2010. p. 1035–1045.

JIAJIA, W. et al. Clustering product features of online reviews based on nonnegative matrix tri-factorizations. In: IEEE. **Data Science in Cyberspace (DSC), IEEE International Conference on**. [S.l.], 2016. p. 199–208.

JIN, W.; HO, H. H.; SRIHARI, R. K. A novel lexicalized hmm-based learning framework for web opinion mining. In: **Proceedings of the 26th annual international conference on machine learning**. Montreal, Quebec: Citeseer, 2009. p. 465–472.

JU, H.; YU, H. Sentiment classification with convolutional neural network using multiple word representations. In: **Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication**. New York, NY, USA: ACM, 2018. (IMCOM '18), p. 9:1–9:7. ISBN 978-1-4503-6385-3. Available from Internet: <http://doi.acm.org/10.1145/3164541.3164610>.

KANG, M.; AHN, J.; LEE, K. Opinion mining using ensemble text hidden markov models for text classification. **Expert Systems with Applications**, Elsevier, v. 94, p. 218–227, 2018.

KHAN, F. H.; QAMAR, U.; BASHIR, S. Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio. **Artificial Intelligence Review**, v. 48, n. 1, p. 113–138, Jun 2017. ISSN 1573-7462. Available from Internet: <https://doi.org/10.1007/s10462-016-9496-4>.

KUSNER, M. et al. From word embeddings to document distances. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2015. p. 957–966.

LEEK, T. et al. The bbn crosslingual topic detection and tracking system. In: CITESEER. **Working Notes of the Third Topic Detection and Tracking Workshop**. [S.l.], 2000.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. **Mining of Massive Datasets**. [S.l.]: Cambridge University Press, 2014. ISBN 9781107077232.

LI, W. et al. Dwwp: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. **Knowledge-Based Systems**, v. 146, p. 203 – 214, 2018. ISSN 0950-7051. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0950705118300558>.

LIU, B. Opinion mining and sentiment analysis. In: **Web data mining: exploring hyperlinks, contents, and usage data**. 2. ed. [S.l.]: Springer Science & Business Media, 2011. chp. 11.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, v. 5, n. 1, p. 1–167, 2012.

LIU, B.; HU, M.; CHENG, J. Opinion observer: analyzing and comparing opinions on the web. In: ACM. **Proceedings of the 14th international conference on World Wide Web**. [S.l.], 2005. p. 342–351.

LIU, L.; LV, Z.; WANG, H. Opinion mining based on feature-level. In: IEEE. **Image and Signal Processing (CISP), 2012 5th International Congress on**. [S.l.], 2012. p. 1596–1600.

MA, S.; ZHANG, C.; HE, D. Document representation methods for clustering bilingual documents. In: **Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology**. Silver Springs, MD, USA: American Society for Information Science, 2016. (ASIST '16), p. 65:1–65:10. Available from Internet: <http://dl.acm.org/citation.cfm?id=3017447.3017512>.

MATHIEU, B.; BESANÇON, R.; FLUHR, C. Multilingual document clusters discovery. In: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. **Coupling approaches, coupling media and coupling languages for information retrieval**. [S.l.], 2004. p. 116–125.

MIHALCEA, R.; BANEA, C.; WIEBE, J. Learning multilingual subjective language via cross-lingual projections. In: **Proceedings of the 45th annual meeting of the association of computational linguistics**. [S.l.: s.n.], 2007. p. 976–983.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: BURGES, C. J. C. et al. (Ed.). **Advances in Neural Information Processing Systems 26**. [S.l.]: Curran Associates, Inc., 2013. p. 3111–3119.

MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.: s.n.], 2013. p. 746–751.

MOGHADDAM, S.; ESTER, M. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In: **Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval**. Beijing, China: ACM, 2011. p. 665–674.

MOHAMMAD, S. M.; SALAMEH, M.; KIRITCHENKO, S. How translation alters sentiment. **Journal of Artificial Intelligence Research**, v. 55, p. 95–130, 2016.

MONTALVO, S. et al. Multilingual document clustering: an heuristic approach based on cognate named entities. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics**. [S.l.], 2006. p. 1145–1152.

MONTALVO, S. et al. Multilingual news document clustering: two algorithms based on cognate named entities. In: SPRINGER. **International Conference on Text, Speech and Dialogue**. [S.l.], 2006. p. 165–172.

NGUYEN, H. T.; NGUYEN, M. L. Multilingual opinion mining on youtube – a convolutional n-gram bilstm word embedding. **Information Processing & Management**, v. 54, n. 3, p. 451 – 462, 2018. ISSN 0306-4573. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0306457317306581>.

ORTIGOSA, A.; MARTÍN, J. M.; CARRO, R. M. Sentiment analysis in facebook and its application to e-learning. **Computers in human behavior**, Elsevier, v. 31, p. 527–541, 2014.

PAVLOPOULOS, J.; ANDROUTSOPOULOS, I. Multi-granular aspect aggregation in aspect-based sentiment analysis. In: **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. p. 78–87.

PESSUTTO, L. R. C.; VARGAS, D. S.; MOREIRA, V. P. Clustering multilingual aspect phrases for sentiment analysis. In: **2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)**. [S.l.: s.n.], 2018. p. 182–189.

PHU, V. N. et al. Fuzzy c-means for english sentiment classification in a distributed system. **Applied Intelligence**, Springer, v. 46, n. 3, p. 717–738, 2017.

PONTIKI, M. et al. Semeval-2016 task 5: Aspect based sentiment analysis. In: **Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)**. [S.l.: s.n.], 2016. p. 19–30.

PORIA, S.; CAMBRIA, E.; GELBUKH, A. Aspect extraction for opinion mining with a deep convolutional neural network. **Knowledge-Based Systems**, v. 108, p. 42 – 49, 2016. ISSN 0950-7051. New Avenues in Knowledge Bases for Natural Language Processing.

PORIA, S. et al. A rule-based approach to aspect extraction from product reviews. In: **Proceedings of the second workshop on natural language processing for social media (SocialNLP)**. [S.l.: s.n.], 2014. p. 28–37.

POULIQUEN, B. et al. Multilingual and cross-lingual news topic tracking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 20th international conference on Computational Linguistics**. [S.l.], 2004. p. 959.

QIU, G. et al. Opinion word expansion and target extraction through double propagation. **Computational Linguistics**, v. 37, n. 1, p. 9–27, 2011.

RAUBER, A.; DITTENBACH, M.; MERKL, D. Towards automatic content-based organization of multilingual digital libraries: An english, french, and german view of the russian information agency novosti news. In: **Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies**. [S.l.: s.n.], 2001.

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. **Knowledge-Based Systems**, v. 89, p. 14 – 46, 2015. ISSN 0950-7051. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0950705115002336>.

RIAZ, S. et al. Opinion mining on large scale data using sentiment analysis and k-means clustering. **Cluster Computing**, Springer, p. 1–16, 2017.

RUDER, S. A survey of cross-lingual embedding models. **CoRR**, abs/1706.04902, 2017. Available from Internet: <http://arxiv.org/abs/1706.04902>.

SAIF, H. et al. Contextual semantics for sentiment analysis of twitter. **Information Processing & Management**, v. 52, n. 1, p. 5 – 19, 2016. ISSN 0306-4573. Emotion and Sentiment in Social and Expressive Media. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0306457315000242>.

SMITH, S. L. et al. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. **CoRR**, abs/1702.03859, 2017. Available from Internet: <http://arxiv.org/abs/1702.03859>.

STEINBACH, M. et al. A comparison of document clustering techniques. In: BOSTON. **KDD workshop on text mining**. [S.l.], 2000. v. 400, n. 1, p. 525–526.

TAN, P. et al. **Introduction to Data Mining**. [S.l.]: Pearson Education, 2013. (What's New in Computer Science Series). ISBN 9780133128901.

TELLEZ, E. S. et al. A simple approach to multilingual polarity classification in twitter. **Pattern Recognition Letters**, v. 94, p. 68 – 74, 2017. ISSN 0167-8655. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0167865517301721>.

TOH, Z.; SU, J. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In: **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 496–501.

TOH, Z.; SU, J. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. San Diego, California: Association for Computational Linguistics, 2016. p. 282–288.

VARGAS, F. A.; PARDO, T. A. S. Aspect clustering methods for sentiment analysis. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 365–374.

WANG, T. et al. Product feature summarization by incorporating domain information. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. [S.l.], 2013. p. 231–243.

WANG, W. et al. Recursive neural conditional random fields for aspect-based sentiment analysis. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. Austin, Texas: Association for Computational Linguistics, 2016. p. 616–626.

WANG, W. et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: **AAAI**. San Francisco, California USA: Association for Computational Linguistics, 2017. p. 3316–3322.

WEI, C.-P.; YANG, C. C.; LIN, C.-M. A latent semantic indexing-based approach to multilingual document clustering. **Decision Support Systems**, Elsevier, v. 45, n. 3, p. 606–620, 2008.

XIANG, B.; ZHOU, L. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. [S.l.: s.n.], 2014. v. 2, p. 434–439.

XIONG, S.; JI, D. Exploiting flexible-constrained k-means clustering with word embedding for aspect-phrase grouping. **Information Sciences**, v. 367-368, p. 689 – 699, 2016. ISSN 0020-0255. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0020025516304832>.

YOGATAMA, D.; TANAKA-ISHII, K. Multilingual spectral clustering using document similarity propagation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2**. [S.l.], 2009. p. 871–879.

YUE, L. et al. A survey of sentiment analysis in social media. **Knowledge and Information Systems**, Jul 2018. ISSN 0219-3116. Available from Internet: <https://doi.org/10.1007/s10115-018-1236-4>.

ZAKI, M. J.; JR., W. M. **Data Mining and Analysis**. [S.l.]: Cambridge University Press, 2014. ISBN 9781107779105.

ZHAI, S.; ZHANG, Z. M. Semisupervised autoencoder for sentiment analysis. In: **AAAI**. [S.l.: s.n.], 2016. p. 1394–1400.

ZHAI, Z. et al. Clustering product features for opinion mining. In: **Proceedings of the Fourth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: ACM, 2011. (WSDM '11), p. 347–354. ISBN 978-1-4503-0493-1. Available from Internet: <http://doi.acm.org/10.1145/1935826.1935884>.

ZHAI, Z. et al. Constrained lda for grouping product features in opinion mining. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2011. p. 448–459.

ZHANG, L.; LIU, B. Aspect and entity extraction for opinion mining. In: **Data mining and knowledge discovery for big data**. [S.l.]: Springer, 2014. p. 1–40.

ZHANG, Y.; LIU, M.; XIA, H.-X. Clustering context-dependent opinion target words in chinese product reviews. **Journal of Computer Science and Technology**, Springer, v. 30, n. 5, p. 1109–1119, 2015.

ZHAO, L. et al. Clustering aspect-related phrases by leveraging sentiment distribution consistency. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1614–1623.

ZHAO, Y.; QIN, B.; LIU, T. Clustering product aspects using two effective aspect relations for opinion mining. In: **Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data**. [S.l.]: Springer, 2014. p. 120–130.