Latent Aspect Detection via Backtranslation Augmentation

Farinam Hemmatizadeh
University of Windsor, Canada
hemmatif@uwindsor.ca

Alice Yu
Vincent Massey Secondary School, Canada
alice.yu@uwindsor.ca

Christine Wong
University of Windsor, Canada
wong93@uwindsor.ca

Hossein Fani
University of Windsor, Canada
histani@uwindsor.ca

ABSTRACT

Within the context of review analytics, aspects are the features of products and services at which customers target their opinions and sentiments. Aspect detection helps product owners and service providers identify shortcomings and prioritize customers’ needs. Existing methods focus on detecting the surface form of an aspect falling short when aspects are latent in reviews, especially in an informal context like in social posts. In this paper, we propose data augmentation via natural language backtranslation to extract latent occurrences of aspects. We presume that backtranslation (1) can reveal latent aspects because they may not be commonly known in the target language and can be generated through backtranslation; (2) augments context-aware synonymous aspects from a target language to the original language, hence addressing the out-of-vocabulary issue; and (3) helps with the semantic disambiguation of polysemous words and collocations. Through our experiments on well-known aspect detection methods across several datasets of restaurant and laptop reviews, we demonstrate that review augmentation via backtranslation yields a steady performance boost in baselines. We further contribute LADy3 (§), a benchmark library to support the reproducibility of our research.

CCS CONCEP'TS

• Computing methodologies → Machine translation; • Information systems → Information extraction.

KEYWORDS

Review analysis; Aspect detection; Backtranslation augmentation;

1 INTRODUCTION

The key characteristic of customers’ opinions is to target aspects of a product or service to convey an opinion. For example, the review “forced us to buy pricey dresses through cheap behavior” entails the aspects ‘dress’ and ‘seller’ toward which opinions of ‘pricey’ and ‘cheap behavior’ is expressed with ‘negative’ sentiments, respectively. Aspect detection is crucial in customers’ review analysis in e-commerce and social platforms [18]; it helps product owners and service providers to identify shortcomings and prioritize improvements according to customers’ needs, hence maintaining revenues and preventing customer churn [1, 13, 32].

Existing methods to detect aspects in reviews and their respective opinions and sentiments have primarily expanded into two subtasks: i) aspect term sentiment analysis which extracts explicit terms that point to an aspect either by tagging the terms [15, 16] or delineating the span of the terms [30, 35] in the review’s text (fine-grained terms), and ii) aspect category sentiment analysis which maps a review into a set of predefined course-grained categories of aspects [12, 24, 27]. In our example, an aspect term extraction method would tag ‘dresses’ as an aspect. However, it falls short of detecting a latent or implicit occurrence of an aspect like ‘seller’. An aspect category detection method would detect the categories ‘clothing’ for ‘dress’ and ‘staff’ for ‘seller’, irrespective of explicit or latent occurrences of aspects.

On the one hand, aspect term sentiment analysis approaches forego the latent occurrences of aspects in reviews, as seen earlier in the example for ‘seller’. Indeed, when writing a review, the customer may write about her opinion and overall rating while overlooking the aspects’ terms for a product or service due to being part of common knowledge. As shown in the literature, 35% of reviews on restaurants and electronics include latent aspects [5, 29]. The need for latent aspects detection is even more pressing in online social review platforms where reviews are unsolicited, short, noisy, and mostly rely on social background knowledge. Furthermore, aspect term extraction methods fall short in the presence of out-of-vocabulary (oov) aspects, i.e., aspects that have been unseen in the training dataset. On the other hand, although aspect category sentiment analysis approaches are agnostic to the occurrence of aspect terms, they heavily rely on occurrences.
of other parts of a review, such as opinion terms whose latency renders aspect category detection challenging.

In this paper, we propose review augmentation via natural language backtranslation to address the latent occurrence and out-of-vocabulary aspects in aspect term extraction methods while overcoming the performance drain caused by the latency of terms in aspect category detection methods. We translate a review from its original language (e.g., English) to a target language (e.g., French), and then translate it back to the original language using a machine translator (e.g., Meta’s nllb [6]). This round-trip translation generates diverse paraphrases of a review while withholding semantic context [33], as a result of which:

1. **Backtranslation can reveal latent aspects as they may not be commonly known in the target language.** For instance, when “forced us to buy pricey dresses by cheap behavior” is translated to Farsi as: “فروشنده با رفتار زندگی ما را جبر به خرید لباس‌های گران می‌کرد”, followed by a backtranslation to English, “with cheap behavior, seller forced us to buy expensive clothes”, it brings up ‘seller’;

2. **Backtranslation can address out-of-vocabulary aspects by augmenting context-aware synonymous aspects from the target language to the original language, as opposed to simple synonym replacement [7, 34].** For example, when “… served me the wrong dish!” is translated to Spanish as “… me sirvió la comida equivocada! …”, followed by a backtranslation to English “… served me the wrong food!”, ’food’ appears for ‘dish’, compared to ‘bowl’ or ‘plate’;

3. **Backtranslation can disambiguate polysemous terms and collocations, leading to the detection of latent aspects.** For instance, translating “… through cheap behavior” to Spanish “… a través de su comportamiento insignificante”, and backtranslating to English “… through her petty behavior” maps the term ‘cheap’ to ‘petty’, which is more semantically related to behavior of a person, leading to the detection of the latent aspect ‘staff’, as opposed to other semantics like ‘inexpensive’ for ‘dresses’.

For similar reasons, backtranslation has been employed in opinion mining [10, 28] and other various natural language processing tasks [9, 17, 23]. However, there has been no study on its synergistic impacts on aspect detection, to the best of our knowledge. In this paper, we systematically benchmark aspect detection models when their training sets of reviews in English are augmented with backtranslated versions through various languages from different language families and study the effects on the performance. We further contribute LADyL (Living Arabic Dataset Library), an extensible and standard benchmark library, to support the reproducibility of our research. Through our experiments on well-known supervised and unsupervised aspect detection baselines across several datasets of restaurant and laptop reviews, we demonstrate that review augmentation via backtranslation has led to a steady performance boost in baselines in different domains.

2 **AUGMENTATION VIA BACKTRANSLATION**

Our goal is to explore the synergistic impact of natural language backtranslation as an augmentation technique on the aspect detection task, especially when aspects are latent.

**Definition 1 (Aspect Detection).** Given a review after normalization of its raw text in a natural language $l$ as a bag (set) of terms $r_p = \{i\}$ about a product or a service $p$ concerning an aspect $a$, which may not be in $r_p$, i.e., $a \not\in r_p$ or $a \in r_p$, the aspect detection aims at identifying aspect $a$ in $r_p$. $A_p$ is the set of all aspects of the product or service $p$ and $R_p$ is the set of all $p$’s reviews.

Our research includes two pipelined components: (1) review backtranslation, further divided into backtranslation and semantic alignment subcomponents, and (2) review aspect detection.

### 2.1 Review Backtranslation

Let $L$ be the set of languages. Given a review $r_p = \{i\} \in L$, we translate it to language $l' \in L$ resulting in review $r_p' \in l'$ and backtranslate it to $l$ which results in $r_p^* = \{i^*\}$. We create an augmented review set $R_p^*$ by adding augmented reviews to the original set $R_p$, which is used for the training phase. The original set $R_p$ can also be augmented with multiple or all languages in $L$.

#### 2.1.1 Backtranslation

In our study, without loss of generality to any machine translation models, we apply Meta’s ‘no language left behind’ (nllb) [6], an open-source neural machine translator capable of providing high-quality translations between 200 languages. We deliberately chose nllb for its particular focus on realizing a universal translation system while prioritizing the needs of underserved communities for low-resource natural languages, as opposed to a small dominant subset of natural languages; it enables review backtranslation augmentation via a vast variety of natural languages with distinct properties. Further, nllb is open-sourced to foster transparency and can be smoothly integrated into any pipeline with few lines of code. For this paper, we translated English reviews into French, German, Spanish, and Farsi from Indo-European language family, Chinese from sino-tibetan language family, and Arabic from Afro-Asiatic language family. We employ the same nllb translator to bring the translated reviews back to English. As shown in Table 1, a backtranslated version of a review may carry term replacement (e.g., ‘food’ for ‘dish’) and/or new terms (e.g., ‘seller’), among other changes.

#### 2.1.2 Semantic Alignment

Although Yu et al. [33] have shown that natural language backtranslation preserves the semantic context, i.e., an original piece of text and its backtranslated versions are about the same topic, we have observed otherwise in review analysis, especially when using languages whose family is different from English’s like Chinese or Arabic. For example, the review “the duck confit is always amazing” became “the duck crib is always fantastic” after backtranslation via Chinese where the aspect ‘duck confit’ (‘food’) has semantically drifted to ‘duck crib’ (‘furniture’). To alleviate semantic drift in backtranslation augmentation, we perform pairwise term alignment [26] between the original review and backtranslated versions to ensure accurate and meaningful machine translation. Term alignment matches the aspect terms of an original review with a corresponding term in the backtranslated review. Moreover, it matches the aspect terms of the original review with its corresponding aligned backtranslated aspect.
Table 2: Statistics on original and backtranslated reviews.

<table>
<thead>
<tr>
<th>dataset</th>
<th>reviews</th>
<th>avg aspect</th>
<th>exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td>semeval-14-laptop</td>
<td>1,488</td>
<td>1.5946</td>
<td>0.1763</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2178</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0272</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3394</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3224</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3702</td>
</tr>
<tr>
<td>semeval-14-restaurant</td>
<td>2,023</td>
<td>1.8284</td>
<td>0.1831</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2236</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2929</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3645</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3724</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4088</td>
</tr>
<tr>
<td>semeval-15-restaurant</td>
<td>833</td>
<td>1.5354</td>
<td>0.2034</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2312</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3021</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3834</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4034</td>
</tr>
<tr>
<td>semeval-16-restaurant</td>
<td>1,234</td>
<td>1.5233</td>
<td>0.2023</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2331</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2991</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.3834</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4034</td>
</tr>
</tbody>
</table>

Figure 1 displays an example where ‘dish’ in the original review has matched with ‘food’ in the Spanish-backtranslated version. For term alignment, we use sinalign-itermax [26], an unsupervised approach to align terms in a pair of parallel texts by calculating the pairwise cosine similarities between the embeddings of words in the source and target texts. We further apply semantic similarity on pairs of review and backtranslated versions using declutr [11] and filtered out backtranslated reviews with less than a threshold (e.g., 0.5) semantic similarity.

2.2 Review Aspect Detection

To assess the contribution of review augmentation using backtranslation, we train two instances of an aspect detection model: one is trained on an original dataset in language $l$, i.e., $R_p$, and another one is trained on augmented datasets via language $l'$, i.e., $R_p'$. We cross-compare the performance of two instances of the model on the same test set with no augmentation.

3 EXPERIMENTS

We seek to answer the following research questions:

RQ1: Does augmentation via natural language backtranslation improve the performance of aspect detection models?

RQ2: Does backtranslation augmentation improve the performance of aspect detection models when aspects are latent?

RQ3: Is the impact of augmentation consistent across datasets from different review domains?

3.1 Setup

3.1.1 Datasets. Our benchmark includes well-known publicly available semeval datasets of reviews in English for aspect detection. We used training sets of semeval-14’s reviews on restaurants and laptops [21], and semeval-15 and semeval-16’s reviews on restaurants [19, 20]. Following the literature [16, 27], we segmented reviews into sentences and performed our experiments on each sentence as an individual review assuming each sentence entails one aspect. Table 2 shows the datasets’ statistics.

3.1.2 Backtranslation and Semantic Alignment. We randomly divided a dataset into 85% training and 15% test sets of reviews. We augmented the training set of reviews through backtranslation using Meta’s nllb [6]. Table 2 shows the average scores of exact match metric between original reviews and their respective backtranslated versions per dataset after filtering out backtranslated reviews with less than 0.5 semantic similarity for semantic drift. As seen, backtranslated versions of reviews from languages that belong to the same family as English, including Spanish, French and German, are more similar compared to Chinese and Arabic. We further merge all backtranslated reviews from all different languages into one set, referred to as all, to explore the impact of backtranslation from all languages at once.

3.1.3 Baselines. We benchmarked the following supervised and unsupervised aspect detection models.

bert-tfm [16] is a supervised tagging-based method that employs contextual embeddings of reviews from bert [8] followed by a self-attention layer.

cat [27] is an unsupervised aspect detection method. It forms a candidate set of aspect terms from nouns. A review is then transformed into an attention matrix using pretrained vectors of constituent terms. Finally, it selects an aspect term that has the highest similarity with the weighted vector of the input review.

loclda [4] is an unsupervised method following the same assumption as in latent Dirichlet allocation [3] where the review’s terms are generated based on an aspect (the review’s topic).

btm [14, 31] is an unsupervised method for short texts like reviews. It learns the aspects by directly modeling the generation of term cooccurrence pairs (biterms) in the entire review dataset to address sparse term occurrences in loclda.

neurallda [25] is an unsupervised topic modeling method based on variational autoencoder, which encodes the term occurrence vector of reviews onto a continuous latent representation as aspects.

cmt [2] extends neurallda with the concatenation of the term occurrence vector and pretrained dense contextual vector representation from sentence-bert [22] in the input layer.

random is a naive method that chooses an arbitrary term as the review’s aspect to provide a minimum base for comparison.

See LADY (§) for the complete list of baselines’ hyperparameters.

3.1.4 Evaluation Methodology. We performed 5-fold cross-validation on a training set for model training and validation, which results in one trained model per fold. Given a review of the test set, we compared the ranked list of predicted aspect terms by the model of each fold with the observed aspect terms and reported the average performance of models in all folds by information retrieval metrics, including normalized discounted cumulative gain (ndcg) as well as classification metrics including precision (pr) and recall (rec). To evaluate how backtranslation augmentation helps with our trained model in detecting latent aspects, given a random review from the test set, we removed the aspect terms from the review (synthetically make it latent) and used our model to predict the review’s latent aspect. We evaluate the baselines on a randomly increasing percentage of latent aspects in the test set from 0% (all aspects are explicit) to 100% (all aspects are latent).

3.2 Results

In response to RQ1, i.e., whether augmentation via language backtranslation improves the performance of aspect detection methods, from Table 3, all baselines, including random, could
generally improve upon integration across datasets and domains. Specifically, the best results were when the training sets were augmented from all languages (+all). On a per-language basis, we can further observe that backtranslation augmentation via languages in the same family as English, such as Spanish, German, or French, generally yielded stronger results. However, there are inconsistencies in backtranslation augmentation across languages that result in different performance gains. For instance, a poor model before augmentation like ctm can boost its performance by augmentation and become one of the best methods among unsupervised baselines. This is particularly true when their training sets were augmented by all languages (+all). Regarding RQ3, i.e., if the impact of augmentation across datasets from different domains is consistent, from Table 3, we see that backtranslation augmentation via all languages (+all) generally improved the performance of models in the test set.

4 CONCLUSION AND FUTURE WORK

We presented an evaluation of backtranslation via multitask learning for the task of aspect detection. Our experiments on backtranslation via six natural languages from varying language families demonstrate the synergistic impact of backtranslation augmentation across aspect detection methods and domains, including restaurant and laptop reviews. Further, future research includes experiments on the impact of augmentation across datasets from different domains and the implications of domain-agnostic synergies of backtranslation augmentation in aspect detection.
REFERENCES


