eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Does Coreference Resolution Improve Aspect-Based Sentiment Analysis?

Author: Rosa-Maria Kristiina Ryhänen

Advisors: Rodrigo Agerri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

Aspect-Based Sentiment Analysis (ABSA) has generally focused on extracting explicit opinion targets and classifying them into polarities and categories. Most approaches ignore implicitly expressed opinions, even though they make up a significant part of language; in fact, approximately 25% of the targets in the SemEval ABSA 2016 English restaurant reviews (Pontiki et al., 2016) are implicit and are not taken into consideration when training a model. We propose to solve a part of the implicit targets with coreference resolution in order to improve two ABSA tasks: opinion target extraction and aspect category detection. Our results suggest that coreference resolution helps to perform opinion target extraction and aspect category detection, when the latter is handled as a multi-label classification task. The data and code are publicly available on GitHub[a].

keywords: Aspect-Based Sentiment Analysis, Coreference Resolution, Opinion Target Extraction, Aspect Category Detection

[a] https://github.com/rosamariaryh/absa-coref

Acknowledgements

I would like to thank Rodrigo Agerri for his guidance, patience and precious feedback during the process, both in person and online, both during the experiments and the writing of the thesis. While the thesis was still in process, I moved from Donostia to start a new job in Barcelona, with which he was supportive, making the writing of the thesis as smooth as it could be.

Thank you Rosa Castrén for being such a great co-worker and for taking the time to comment my thesis in your free time. Your linguist insight was much appreciated and I know I can count on you for feedback and interesting conversations in the future as well.

Lastly, I want to thank everyone who has supported me during the thesis or showed interest in it. Family, friends, colleagues at TrustYou and M47 Labs, flatmates and classmates: thank you for being there for the entire time.

Contents

List of Figures

List of Tables

1 Introduction

In the recent years, sharing opinions on online platforms and social media has increased thanks to technological advances. Individuals and organisations can turn to reviews for decision-making. Reviews are one of the most common ways to share opinions, and they are available for all types of products and services on online shops, but also on websites and applications geared towards reviews, like Google Maps, Yelp, Trip Advisor and Trustpilot.

From a customer's point of view, it is beneficial to know other people's opinions before purchasing a product or hiring a service in order to ensure quality. In fact, in 2021, 77% of the customers read reviews either regularly or always before making a decision on the purchase (BrightLocal, 2022). At the same time, people have shifted from relying on experts to searching for information online regarding their big decisions in life in the US (PewResearch, 2020), which also explains the heightened importance of online presence. Considering this, automatic opinion extraction and review summaries may alleviate the burden of scrolling through all reviews.

Online presence has a great impact on sales from a company's point of view as well. Having positive reviews attracts new customers and replying to negative reviews may help to maintain a good brand image. Doing so can also help to appear higher in the search engines and increase web traffic. However, it is time-consuming to read all reviews, which is why a sentiment classifier may help to detect the negative reviews and an opinion target extraction model can give important insight on the most mentioned aspects of the service or product.

1.1 Aspect-Based Sentiment Analysis

As stated above, with the amount of data keeps growing, it becomes unfeasible to manually review all the texts to get a more global opinion of the product. Thanks to improved machine learning systems, it is possible to process texts and extract valuable information from them. This is exactly the case of aspect-based sentiment analysis, which is one of the main ways to convert polarised, unstructured, raw texts into meaningful knowledge. Such processes of conversion from data to knowledge include: extracting opinion targets and clustering them to analyse the most mentioned items in a text (Figure 1), classifying opinion targets into negative and positive sentiments to discover the best and worst rated aspects of a product or service, and grouping the opinion targets into broader categories for easier data processing (Figure 2). A good example of this is hotel reviews, where a category may be *for couples* and the most frequent opinion targets may be *king-sized bed, roses, candles* and *tranquility.*

However, not all types of opinions can be directly processed because language naturally includes a great number of context-dependent pragmatic elements, such as presuppositions (Karttunen, 1974; Ducrot, 1969 and implicatures (Over and Grice, 1989), which require reasoning and world knowledge. Another example is discourse referents Karttunen (1976), which are easily understood by humans, but complicated to process for Natural Language Processing (NLP) systems; for example, it is not possible to extract an opinion target and

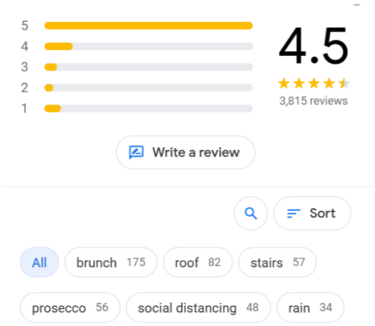


Figure 1: Opinion Targets and Their Frequency for a Bar on Google Maps

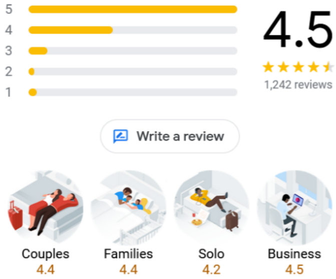
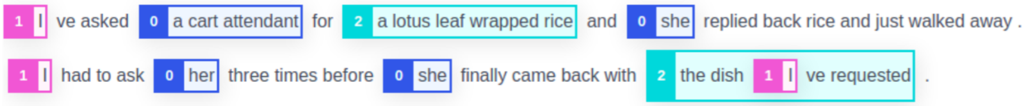


Figure 2: Opinion Categories for a Hotel Based on Opinions on Google Maps

classify it into a category if the target is a pronoun, as it is semantically empty. When opinions are analysed by automatic means, they must be explicit, which means that any ambiguous words or implicit expressions are simply ignored. This naturally has an impact on the quality of annotations in the data set and, consequently, on the the model that we train.



1 I ve asked 0 a cart attendant for 2 a lotus leaf wrapped rice and 0 she replied back rice and just walked away .

1 I had to ask 0 her three times before 0 she finally came back with 2 the dish 1 I ve requested .

Figure 3: Example of Coreference with Three Entities

One way to increase explicit information in the text is to exploit references, as is the case of coreference resolution, seen in Figure 3. In this example, we observe a coreference cluster for food (2) *a lotus leaf wrapped rice* and *the dish I've requested* and for service (0) *a cart attendant*, *her* and *she*. Coreference resolution establishes links between the words referring to the same entity and consequently replaces the implicit references with an explicit one. This way, the information is made explicit and can be used in different NLP tasks, such as in sentiment analysis, summarisation and question answering (Sukthanker et al., 2020).

While both aspect-based sentiment analysis and coreference have quite an extensive history of research separately, little research has been carried out on coreferential models applied to ABSA tasks; even more so for neural coreferential models, as few have been used to resolve implicit targets in sentiment analysis.

1.2 Research questions

The motivation for this thesis stems from the lack of research on implicit targets in several ABSA tasks. They have been noted ever since the first ABSA investigations (Hu and Liu, 2004b), but have generally been disregarded in research; Section 2.3 gives an overview of this particular issue. Thus, the novelty of the thesis lies in applying coreference resolution on implicit opinion targets in order to observe its effects on ABSA tasks.

We investigate whether opinion target extraction and aspect category detection can be improved by explicitising implicit language with coreference resolution methods. We assume that training models with a greater number of explicit targets will help us perform the aforementioned ABSA tasks better. Our main hypotheses are the following:

H1 Coreference resolution helps to extract opinion targets in texts

H2 Coreference resolution helps to detect aspect categories of the opinion targets

The objective of this thesis is to discover whether coreference resolution helps to perform some of the most common ABSA tasks: opinion target extraction and aspect category detection. Our contributions are the following:

- We address implicit coreferential targets with language models, which has not gained enough attention in ABSA research.
- We analyse quantitatively and qualitatively the implicit targets of ABSA SemEval 2016 English restaurant reviews.
- We analyse the limits and common errors of the AllenNLP coreference model.
- We manually annotate the aforementioned reviews with coreference resolution [1].
- We automatically annotate the aforementioned reviews with coreference resolution[2].
- We treat opinion target extraction as a sequence labelling task, in which we conclude that coreference resolution is beneficial for OTE.

[1] Publicly available at `https://github.com/rosamariaryh/absa-coref`

[2] Publicly available at `https://github.com/rosamariaryh/absa-coref`

- We handle opinion target extraction and classification, a more complex task, with sequence labelling methods and conclude that manual coreference resolution improves the results.
- We formulate aspect category detection as a multi-class classification task where coreference resolution has a minimal effect on the results.
- We formulate aspect category detection as a multi-label classification task and conclude that manual coreference improves the system.
- We show that having corpora formed at a document level is beneficial for tasks related to coreference resolution, which require this type of formatting.
- Lastly, as opposed to previous methods, we demonstrate that deep-learning based methods for automatic coreference resolution help to perform ACD and OTE.

In order to test our hypotheses, the thesis is organised as follows: in the next section, we review the state of the art for ABSA tasks and coreference resolution. In Section 3, we explain the algorithms and data set that were used in this thesis. Next, in Section 4, we discuss the experimental setting, which justifies the need for this type of work and details the steps taken to test our hypotheses. We present the results in Section 5, which is followed by an error analysis in Section 6. Finally, we conclude the thesis in Section 7 and offer prospects for future research in Section 8.

2 Related work

This section aims to give an overview of the state of the art of aspect-based sentiment analysis, coreference resolution and coreference resolution methods applied to ABSA. We start by introducing sentiment analysis, continue with explaining the subtasks, shared tasks and experiments carried out in ABSA, review implicit language in the context of sentiment analysis, and, finally, explain the coreference resolution techniques and their results.

2.1 Introduction to Sentiment Analysis

Sentiment analysis is the field of study in NLP that studies people's attitudes and opinions on a vast variety of texts, often including but not limited to reviews, tweets, comments, and blog posts. Although related topics like subjectivity analysis (Wiebe et al., 1999) or semantic analysis (Hatzivassiloglou and McKeown, 1997) sparked curiosity in researchers in the 90s, it was not until the turn of the millennium that the first works of sentiment analysis were carried out. This is greatly due to technological advances like the expansion of the Internet and, thus, to the amount of digitised data that was available (Liu, 2012). With the rise of online commerce in the early 2000s, it was discovered that customer reviews play a significant role on the purchase power, which is why there was a great motivation to process the data and turn it into meaningful information.

Sentiment analysis is a polarity classification task that is often performed on a document-level (Hu and Liu, 2004b; Pang and Lee, 2008; Liu, 2012). This means that given an input text, we obtain an output label for its class, such as *positive*, *negative* or *neutral*, which determines the overall sentiment of a text. Similarly, we can classify texts according to their category, such as *food*, *service* or *ambience*. This method is efficient, but ignores any differences we may encounter in the text.

Sentence-level polarity classification can be carried out in order to get more truthful insight of the sentiment in the text. This method allows us to measure the strength of sentiments in a text that contains both negative and positive sentences, assuming that there are not multiple opinions in one sentence.

However, neither of the aforementioned approaches considers that a single sentence may have several opinions with the same polarity or that the polarities may be linked to different targets, such as in example (1), where we observe that *call quality* is positive, but *battery life* is negative.

1. The iPhone's call quality is good, but its battery life is short. (Liu, 2012)

Not only does sentence-level sentiment analysis ignore multiple polarities in a sentence, but it also dismisses the origin of the polarities, which are the opinion targets. It is not sufficient to say that example (1) contains a positive and negative opinion, especially if these are reduced to an overall score, which would result in a neutral opinion.

In this context, a paradigm shift occurred and the necessity for Aspect-Based Sentiment Analysis emerged. In contrast to sentiment analysis, ABSA is a more refined version, which seeks to identify the targets of a text and classify their category and polarity. Some of the first studies for ABSA were carried out for customer reviews in order to facilitate customers' decision making for online purchases by the help of summarisation (Hu and Liu, 2004a,b). This means that, as Liu (2012) explains, "[ABSA] turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses." The following sections outline the main tasks and data used in ABSA.

2.2 Aspect-Based Sentiment Analysis

As previously mentioned, the first works of ABSA emerged in the early 2000s and include systems based on statistics and rules (Hu and Liu, 2004b,a; Pang and Lee, 2008). Hu and Liu (2004b) used the NLProcessor[3] to perform parsing, POS-tagging and chunking, and used association mining to identify the most common aspects of a product, which they refer to as *features*. Then, pruning was applied to remove a part of the candidate features generated by the association mining algorithm in order to avoid overfitting and to make the model more efficient. Additionally, a WordNet[4]-based system was created to predict polarities in sentences and classify them.

Both supervised and unsupervised methods have been used in ABSA tasks. Popular supervised systems include feature engineering and algorithms like the Support Vector Machine or Conditional Random Fields (Liu, 2012; Wagner et al., 2014; Barnes et al., 2018). The SemEval ABSA experiments use SVMs and are discussed in Section 2.2.3. These offer robust solutions for ABSA, but the CRF is computationally highly complex, which makes it more difficult to retrain models. On the other hand, SVMs take long to train, which is why they are not recommended for large data sets, and choosing the correct kernel is a complicated task. Also, interestingly, Wagner et al. (2014) discovered that their SVM system with bag-of-words and sentiment lexicon features did not perform much better than their rule-based system.

Weakly supervised methods have also been used for ABSA tasks. These include methods using topic modelling, more specifically Latent Dirichlet Allocation (Blei et al., 2003). García-Pablos et al. (2017) created an almost unsupervised system that takes as input unlabelled text and a seed word for one category, and consequently outputs aspect-terms, positive words and negative words. Their approach combines topic modelling with continuous word embeddings and a Maximum Entropy classifier, which can be used for various languages and domains (García-Pablos et al., 2017). The advantage of an LDA model is its domain-adaptability and the fact that no labelled data is needed, which can save a significant amount of time and effort. However, syntactic and semantic information is generally ignored and it may be difficult to evaluate their performance.

Pre-trained language models, like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have alleviated the problem of feature engineering and give the flexibility to adapt models for different ABSA tasks. In this context, many

[3] http://www.infogistics.com/textanalysis.html

[4] http://wordnetweb.princeton.edu/perl/webwn

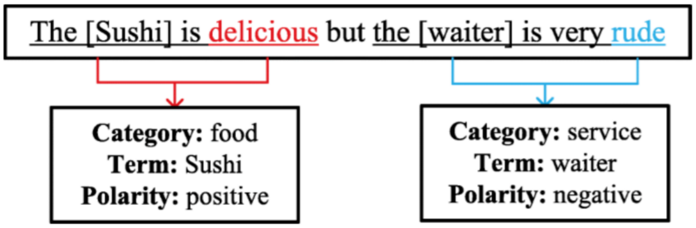


Figure 4: Aspect-based sentiment analysis (Lu et al., 2021)

experiments have been carried out with the SemEval datasets. Hoang et al. (2019) fine-tuned a BERT-based model to find relations between targets and texts and sentiment contexts, both for in domain and out-of-domain aspects. Additionally, Sun et al. (2019) formulated this as a sentence pair classification task, using question answering and natural language inference techniques to improve the system, and Li et al. (2019b) explored the BERT embedding component for End-to-End Aspect-Based Sentiment Analysis, which showed great robustness and effectiveness as opposed to BERT-based models.

2.2.1 Subtasks of ABSA

Generally, Aspect-Based Sentiment Analysis is treated as three subtasks: opinion extraction, category classification, and polarity classification. Although it is common to only extract the opinion targets, it may be of interest to extract the polarised sentiment terms as well so that maximum information is obtained, or perform the tasks as a pair or triplet extraction, in which case we could retain the relation between each opinion target and polarised word (Li et al., 2019a; Wang et al., 2017; He et al., 2019). That being said, ABSA is usually performed in the following steps:

- Opinion Target Extraction (OTE) intends to identify the aspects or targets of the text
- Aspect Category Detection (ACD) classifies the opinion targets into categories
- Polarity detection classifies the opinions according to their polarity, often *positive*, *negative* or *neutral*

Opinion target extraction (OTE) is often treated as a token-level sequence labelling task. One of the first attempts at OTE was by Hu and Liu (2004b) who used an unsupervised algorithm relying on POS-tagging, feature extraction and association mining for extracting frequent items. Other early studies were carried out by Kim and Hovy (2006), where semantic role labelling was used and by Zhuang et al. (2006), who also presented a system based on POS-tagging.

Poria et al. (2016) was one of the first to apply deep learning approaches for OTE. A 7-layer convolutional neural network that tags each word as a target or non-target word was used together with a system of heuristic linguistic patterns. Doing so significantly improved the system, as they discovered that a non-linear model better fits the data than linear models, like CRFs or SVMs. Additionally, the pre-trained word embedding features helped to outperform other methods (Poria et al., 2016). Another approach includes that of Li and Lam (2017), who introduced an LSTM-based system, which outperformed the best systems in the SemEval task by 3 points. Li et al. (2018) further improved the LSTM-based system with two key components: Truncated History-Attention (THA) for aspect detection history and Selective Transformation Network for opinion summary. Doing so further improved the results, outperforming several CRF and LSTM-based systems.

Agerri and Rigau (2019) created a language-independent system to extract opinions in six languages, obtaining significantly better results compared to previous papers. They used a sequence labeller from the IXA pipes[5], which relies on the perceptron algorithm, and combines and stacks language-independent word representation features. Creating a system that does not use linguistic features for training is extremely beneficial for transferring the model to other languages and for reducing human intervention, and doing so resulted in an improvement of around 6 to 7 points in the F1 score for the English SemEval data from 2014 to 2016.

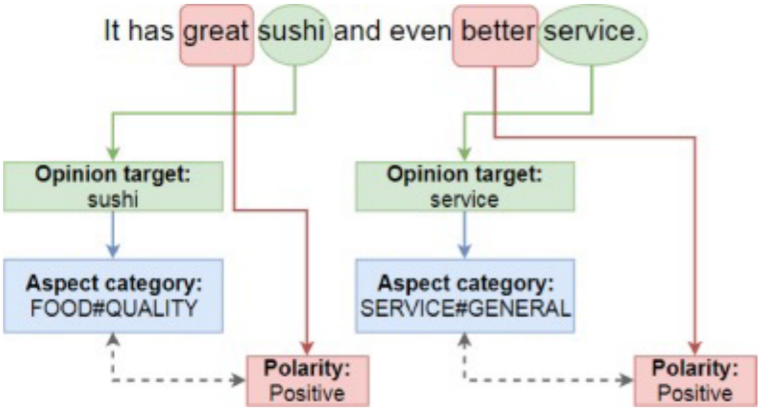

It has great sushi and even better service.
Opinion target: sushi
Aspect category: FOOD#QUALITY
Polarity: Positive
Opinion target: service
Aspect category: SERVICE#GENERAL
Polarity: Positive

Figure 5: OTE, ACD, and polarity extraction for ABSA (Do et al., 2019)

Aspect category detection (ACD) has usually been approached as a multi-label text classification task, which has been solved with rules, unsupervised and supervised methods. Liu (2012) presents the main strategies to perform this task:

1. Extraction based on frequent nouns and noun phrases
2. Extraction by exploiting opinion and target relations
3. Extraction using supervised learning
4. Extraction using topic modelling

[5] https://github.com/ixa-ehu/ixa-pipe-opinion

With the rise of transformers, systems utilising attention have grown popular in ACD tasks. Wang et al. (2016) presented a novel approach that includes an LSTM-based model with attention, whereas He et al. (2017) proposed the first unsupervised attention-based model for ACD, exploiting word co-occurrence patterns through neural word embeddings. This gave significantly better results than previous unsupervised methods for ACD, like LDA-based models (Su et al., 2008; Mukherjee and Liu, 2012), due to its ability to capture contexts. Similarly, Ramezani et al. (2020) used contextual representations in their BERT and multi-layer perceptron-based model to detect aspects. Their approach resulted in improvements compared to several supervised and unsupervised baselines and, although unsupervised models have the advantage of not using annotations, Ramezani et al. (2020) argue that ”unsupervised models cannot learn semantic features of the domain very well”, which suggests that supervised methods may be more suitable for ACD.

Seeing that supervised methods have generally given better results, but the cost of data annotation is high, other approaches have been studied. These include few-shot learning, as it allows to train models with less data (Hu et al., 2021; Zhong et al., 2021) and prompting (Min et al., 2021b; Liu et al., 2021; Schick and Schütze, 2021), which formulates ACD as a text generation task and exploits a pre-trained language model's masked language model, obtaining superior results.

Lastly, although opinion target extraction and aspect category detection are treated as two separate tasks, some approaches join them (Li et al., 2019a; He et al., 2019; Wang et al., 2017; Yan et al., 2021. This can decrease computational complexity and increase accuracy, as the possible errors are not propagated from the first task (OTE) to the second task (ACD) (Wang et al., 2017). However, if dependency parsers do not accurately extract the relations between opinion terms and aspect terms, this can lead the system to poorly functioning systems.

2.2.2 SemEval ABSA Data

The most well-known benchmark data sets for resolving aspect-based sentiment analysis have been obtained from the ABSA SemEval workshops, which include several subtasks, attracting various submissions from different teams in every edition. Up to date, four editions (2014, 2015, 2016) have been published and, more recently, a structured sentiment analysis task has been finished[6]. It is important to note that what we refer to as opinion targets in this thesis are called aspect terms in the early SemEval ABSA tasks.

The first shared ABSA SemEval task was Task 4[7] in 2014 and it included annotated datasets for restaurants and laptops (Pontiki et al., 2014). The task was to extract aspect terms and their polarity, as well as the categories of the targets and their polarity. Thus, two evaluation tasks were organised: phase A for aspect terms and coarse categories, and phase B for aspect term polarity and aspect category polarity. The aspect category is generally based on meronymic relationships, such as *food* for category if *chicken* is the

[6] https://jerbarnes.github.io/downloads/SemEval$_2022_Task_10.pdfs$

[7] `https://alt.cri.org/semeval2014/task4/`

aspect. The tasks were performed on a sentence-level, and their context is not taken into account.

The task was further elaborated in 2015[8] by redefining the aspect category as a combination of an entity and attribute, using the form ENTITY#ATTRIBUTE, and relying on a list of entities and attributes that can be combined. The aspects, categories and their polarities were extracted similarly to Task 4 in 2014 but, additionally, the domains were extended to hotels, and out-of-scope sentences were handled (Pontiki et al., 2015). Also, a unified framework and sentence-level tuple were created for the opinions (aspects, opinion target expressions, sentiment polarity) thanks to the established guidelines. Finally, evaluations for out-of-domain data were considered for this task.

In 2016, ABSA was presented as SemEval Task 5[9] and it included texts for 7 domains and 8 languages (Pontiki et al., 2016), as well as common guidelines and an evaluation procedure, which meant participating teams used the same system to evaluate the performance. Both sentence-level and document-level annotations were performed, depending on the language and domain.

The latest ABSA SemEval Task was presented in 2022 as a structured sentiment analysis task, where not only opinion targets, expressions and polarities are extracted, but also the opinion holder is included. Seven datasets were created in five languages: English, Spanish, Catalan, Basque and Norwegian and the domains contain news, hotel reviews, online university reviews and other varied reviews (Barnes et al., 2022).

There are other data sets based on the SemEval ABSA tasks worth noting, such as TOWE (Target-oriented Opinion Words Extraction) and MAMS (Multi-Aspect Multi-Sentiment)[10]. TOWE includes extended annotations for the English restaurant reviews from 2014, 2015 and 2015, as well as for the 2014 laptop reviews with annotated opinion words and annotated relations between the opinion target and opinion word. In the original SemEval data sets, only the opinion targets are annotated, and the polarised opinion words are left untouched. The authors claim that this type of annotations may be beneficial for pair-wise opinion summarisation.

MAMS, on the other hand, provides is a data set consisting of sentences that include at least two opinion targets with differing sentiment polarities. Two versions are available: one for OTE and another for ACD. Using a data set like MAMS could help the polarity classification task focus on classifying real aspects instead of classifying the whole sentence. It was verified that, judging by the results of the system with the SemEval 2014 restaurant data set MAMS is a more complex task than original ABSA tasks (Jiang et al., 2019). For this kind of tasks, the authors state that attention-based models lose word order information perform extremely poorly on MAMS, as the model is not able to identify the context of the aspect.

[8] `https://alt.qcri.org/semeval2015/task12/`

[9] `https://alt.qcri.org/semeval2016/task5/`

[10] Both data sets available at `https://github.com/jiangqn/Aspect-Based-Sentiment-Analysis`

2.2.3 SemEval ABSA Experiments

As mentioned above, the ABSA shared tasks have been organised into various subtasks and several teams submitted their results with the same data. For this thesis, the relevant tasks were opinion target extraction and sentence-level aspect category detection for English restaurant reviews, which are briefly reviewed in this section.

The SemEval ABSA subtasks for OTE use several machine learning systems. In 2014, the best results for opinion target extraction were obtained by DLIREC by using Conditional Random Fields together with POS-tagging and dependency tree features (Pontiki et al., 2014). Their system also used additional clustered features from Yelp and Amazon reviews. Moreover, the tokens of an opinion target were annotated only if they were present in a dictionary that contains all the opinion targets of the training sentences. The target slot was filled with the first target occurrence identified in the sentence and if no target was found, the value NULL was assigned. Doing so resulted in an F1-score of 84.01 for English restaurant reviews.

In 2015, the best submission was by the EliXa team and it achieved an F1 score of 70.05, which was obtained by using an averaged perceptron together with the BIO tagging scheme (Pontiki et al., 2015). Their features included n-grams, token classes, n-gram prefixes and suffixes, and word clusters learnt from Yelp for Brown and Clark clusters and from Wikipedia for word2vec clusters. As in 2014, lists of OTEs from the training data were created, but this time separately for each aspect category.

Finally, in 2016, all participating teams used the same baseline system for OTE Pontiki et al. (2016). Lists of opinion targets were extracted from the training data as opinion target and aspect category tuples, which were used on the test sentences to identify the targets together with the category. The best results for this task was an F1-score of 52.607. Opinion targets were also extracted without their aspect category, for which the best result was an F1-score of 72.34

For the ACD tasks, the best results in 2014 were obtained by NRC-Canada with a system comprising of five binary Support Vector Machines with features based on stemmed n-grams, parse trees, and sentiment lexica learnt from Yelp data (Pontiki et al., 2014). The authors reported that especially using the lexica improved the results greatly, which gave an F1-score of 88.57.

In 2015, the methods used were similar to that of 2014; the system that obtained the best results was created by NLANGP and handled ACD as a multi-class classification problem and exploited features based on 1000 most common unigrams excluding stop words, as well as relying on parsing and word clusters learnt from Amazon and Yelp data (Pontiki et al., 2016). The best results gave an F1-score of 62.68. In 2016, the participating teams trained a Support Vector Machine with a linear kernel similar to 2015 Pontiki et al. (2016). Feature vectors were built for test sentences, after which probabilities were assigned with a category label, using a threshold of 0.2 in the decision-making. This way, an F1-score of 73.031 was obtained.

Moreover, Pontiki et al. (2016) states that there is a trend that focuses on ACD rather than on OTE in ABSA recently: ”An interesting observation is that, unlike SEABSA15, Slot1 (aspect category detection) attracted significantly more submissions than Slot2 (OTE extraction); this may indicate a shift towards concept-level approaches". This means that any opinion target related tasks have gained less importance, although they are fundamental analysing texts thoroughly. Also, if we only consider coreferential implicit language in ABSA, we find even less mentions about it, although the phenomenon is common. The next section aims to introduce the research that has been conducted on coreferential implicit targets.

In 2022, the participating teams were given two baselines: a dependency graph prediction model and a sequence labelling pipeline (Barnes et al., 2022). With them, monolingual structured sentiments and cross-lingual structured sentiments were extracted with data outlined in Section 2.2.2. The best results for the English monolingual system were by ZHIXIAOBAO, and they used the dependency graph prediction model and RoBERTA-large, added an attention mechanism to detect spans, and exploited BERT's masking system to learn suffixes. They discovered that removing the LSTM layer from the original model helped to improve the results.

2.3 Implicit Targets in Sentiment Analysis

To our knowledge, little research has been conducted on the effect of coreference resolution on opinion target extraction and nearly none for aspect category detection. Early work has been presented for opinion mining in general, focusing on polarity extraction on a sentence or document level. Moreover, the majority of the systems used in the following papers use traditional machine learning models or rules and few state-of-the-art models are mentioned.

It should be noted that automatic coreference systems generally focus on the resolution of noun phrases and pronouns, which means that references to previous phrases are ignored, even though these may contain significative events. An example of this in the SemEval ABSA 2016 data set is *I complete the total bagel experience by having it lightly toasted.* which is followed by *Murray won't do it*, where *do it* refers to *having it lightly toasted.* These references are not processed in any way.

Although numerous experiments have been carried out with the ABSA SemEval data, most approaches ignore implicit language. In 2014, it was noted that for the laptop domain, many authors referred to the target implicitly through pronouns and adjectives (e.g., 'expensive', 'heavy'), rather than using explicit terms (e.g., 'cost', 'weight'). In these cases, it was instructed to tag only explicit aspect terms, leaving the referring adjectives unannotated. The same guideline was adopted for ABSA SemEval 2015 and 2016. We observe that the implicit targets are simply annotated as "NULL" and left unprocessed in all editions.

Out of all the data available for SemEval ABSA tasks, only the English restaurant data contained a significant number of implicit targets which could be resolved via coreference resolution. This is mainly due the fact that this particular dataset, unlike the rest, includes full documents instead of isolated sentences. This is also true for datasets used in the SemEval 2022 structured sentiment analysis task, like for OpeNER based dataset and

Multibooked.

2.3.1 Coreference for Polarity Classification

Some of the first approaches were applied on sentence-level sentiment analysis without further extracting targets and their corresponding categories or sentiments. These systems took advantage of linguistic information, such as dependency parsing, chunking and ontologies, and use features together with traditional machine learning methods.

Nicolov et al. (2008) applied coreference resolution to blog texts and studied its effect on sentence polarity classification. They apply a proximity-based algorithm which used extended context windows around nominal and pronominal elements, and calculated polarity scores for the sentences with the help of WordNet[11] relations. This increased the performance for polarity classification by 10%, which means that explicitising coreferential pronouns is beneficial for this ABSA task.

Hendrickx and Hoste (2009) also studied the effect of applying automatic coreference resolution on blog posts, but also on news comments and more formal newspaper articles. Their system uses memory-based machine learning and is based on previous work of Hoste (2005) for Dutch, treating coreference as a binary classification task in which noun phrase relations are studied and a feature vector for each relation is created. The results were significantly better for newspaper articles than for blog posts or comments, which indicates that resolving references in an informal context is more complex. Furthermore, their error analysis revealed that coreferential links requiring world knowledge were difficult to resolve, and that some multi-token noun phrases were incompletely detected.

2.3.2 Coreference for OTE

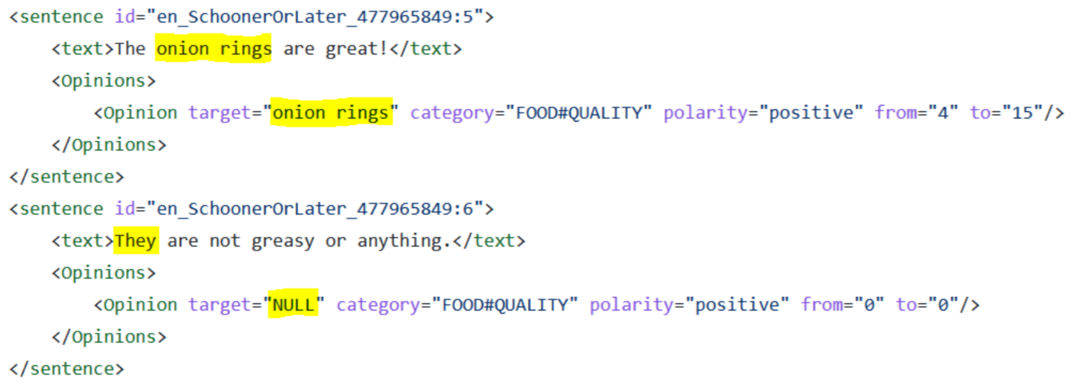
Ding and Liu (2010) applied coreference resolution on forum discussions related to technology and cars with supervised methods. Both opinion targets and attributes were manually annotated and the J48 decision tree was used on WEKA[12]. These results were compared to two baselines: a decision tree and a centering theory based on semantic information. Also, new features like sentiment consistency, mining comparative and entity and opinion word association were added, They discovered that coreference resolution and the new features improved the results around 9 points for baseline 1 and approximately 5 points for baseline 2.

Mai and Zhang (2020) studied the impact of coreference resolution on opinion target extraction with unsupervised learning. A rule-based approach was applied due to its independence of annotated data; more specifically, noun chunking via dependency parsing was used for aspect extraction, after which coreference resolution was applied, testing whether each noun chunk refers to an existing aspect and calculating its cosine similarity. This method improved their previous system's F1 score by 8%, from 0.7 to 0.78, which means that coreference resolution improved the extraction of opinions. It is equally interesting

[11] https://wordnet.princeton.edu/

[12] https://www.cs.waikato.ac.nz/ml/weka/

to observe the great difference between precision and recall: the precision of the opinions extracted improved a staggering 19%, whereas in terms of recall, the results improved only 5%.



```
<sentence id="en_SchoonerOrLater_477965849:5">
    <text>The onion rings are great!</text>
    <Opinions>
        <Opinion target="onion rings" category="FOOD#QUALITY" polarity="positive" from="4" to="15"/>
    </Opinions>
</sentence>
<sentence id="en_SchoonerOrLater_477965849:6">
    <text>They are not greasy or anything.</text>
    <Opinions>
        <Opinion target="NULL" category="FOOD#QUALITY" polarity="positive" from="0" to="0"/>
    </Opinions>
</sentence>
```

Figure 6: Coreferential targets in the SemEval ABSA 2016 data set

2.3.3 Coreference for ACD

Clercq and Hoste (2020) studied the effect of applying coreference resolution on aspect category detection by using the English SemEval ABSA 2015 (Pontiki et al., 2015) restaurant reviews and on the Dutch SemEval ABSA 2016 (Pontiki et al., 2016) restaurant reviews using both manually resolved and automatic links. The COREA system (De Clercq et al., 2011) was used for Dutch, whereas the deterministic Standford Coreference Resolver (Lee et al., 2013) was used for English. Their approach included a bag-of-words method with features from WordNet and DBpedia, as well as classification with a Support Vector Machine. They also took advantage of the annotated category of the NULL target in their work, establishing a semantic category for the implicit target, but this proved to have an insignificant impact. The best results for both languages were obtained with the gold coreference links, an improvement of 0.97% for English and 0.43% for Dutch, leaving room for great improvement.

2.4 Coreference Resolution

Coreference resolution is an NLP task that seeks to identify mentions of an entity and cluster them together (Jurafsky, 2000). In this context, we find anaphoras (referring elements with an antecedent mention) and cataphoras (referring elements with a postcedent mention). The referring elements are typically noun phrases, synonyms or pronouns, which are used for stylistic reasons in order to avoid repetition and create coherence in texts. These are easily understood by humans, but for machines, this task is significantly more complex if several references are present. Some popular approaches for coreference resolution include parsing and semantic information, but deep learning methods have been proven to be more efficient.

The first end-to-end coreference resolution model without syntactic parsing or other external resources was presented by Lee et al. (2017). The novelty of this method lies in treating all token spans as possible coreference mentions and in computing span embeddings, which have context-dependent span boundary representations and cluster heads calculated with an attention mechanism. This way, all span representations are given a score and the unlikely spans are pruned away. Their ensembled ELMo-based model gained an improvement of 3.1% in the F1 score and their single model a 1.5% improvement, obtaining the scores 68.8 and 67.2, respectively.

Subsequently, Lee et al. (2018) presented a system that, apart from using the aforementioned span ranking, included two new features: higher-order inference or, in other words, interactions between the spans, and an additional coarse-to-fine pruning step. Higher-order inference exploits attention iteratively from antecedent spans to make decisions on future spans. These approaches further improved the accuracy on the English OntoNotes benchmark while creating a more computationally efficient system. This gave promising results; nearly 6% of improvement in the F1 score for the models using higher-order inference or coarse-to-fine pruning.

Joshi et al. (2019) extend the end-to-end system by replacing the LSTM-based encoder with a BERT encoder that fine-tunes the model. They presented two ways of improving the system: with individual segments up to 512 tokens and with overlapping segments, which, in a way, exceed the token limit of 512. However, using overlapping segments did not provide any improvement, which indicates that larger context windows for pretraining is possibly not beneficial. With the individual segments, their BERT-large model improved over the ELMo-based model (Lee et al., 2017) 3.9% on OntoNotes and 11.5% on GAP.

Bert for coreference resolution was further improved by creating SpanBERT (Joshi et al., 2020). Seeing that the system did not gain efficiency from overlapping segments, the individual variants were used to train the model. SpanBERT masks random spans instead of individual tokens, and optimises the spans relying on a novel span-boundary objective, the only information used to predict span boundary representations. This substantially improved the representations of spans and, thus, the coreference systems due to the understanding of complex noun phrases. This obtained an F1-score of 79.6, 2.5 data points better than Google BERT and over 6 points better than Lee et al. (2018).

In the latest research, coreference resolution has been formulated as a question answering task (Wu et al., 2020). Generating a query for each mention leverages two problems in previous coreference resolution models: the connection between mentions and their contexts is not considered because the scoring is performed on the output layer, giving somewhat superficial results in terms of context capturing. On the other hand, some mentions can be left out, because the model only works with the proposed mentions. This is true for even the state-of-the-art models, as stated by Zhang et al. (2018). Wu et al. (2020)'s system improved the CoNLL-2012 benchmark by +3.5%, giving an F1 score of 83.1.

3 Methodology

This section explains the material used in the experiments and their evaluation. More specifically, these include data, systems, and evaluation methods.

3.1 Multilingual Language Models

The systems used in our experiments rely on Multilingual Language Models (MLM) based on the Transformer architecture, the state of the art deep learning architecture for various NLP tasks. Transformers are based on an attention mechanism without relying on previous neural architectures, like convolution or recurrent networks, which results in the ability to parallelise predictions (Vaswani et al., 2017). This has been proved to be an effective way to resolve long-distance dependencies between elements and train models quicker, which were a problem in previous approaches like RNNs.

Transformers consist of an encoder and decoder, as seen in Figure 7, and can be used either separately or together when creating transformer-based models. The encoder is generally related to language understanding tasks, whereas the decoder is associated with language generation tasks, and, as mentioned, tasks like machine translation or conversational systems use both. The transformers are easily available online for use via HuggingFace[13]

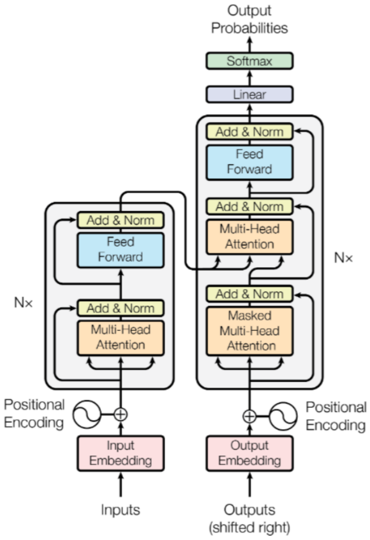


Figure 7: Transformer architecture: encoder and decoder (Vaswani et al., 2017)

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model consisting of 12 or 24 layers of encoders (Devlin et al., 2019). It has two parts: the pre-trained language model and a task-specific fine-tuned layer. The

[13] `https://huggingface.co/models`

language model was trained with unlabelled data with two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). Around 2.5GB of data from Wikipedia and 0.8GB of data from BooksCorpus was used for the training.

Masked language modelling consists of masking approximately 15% the tokens with either an incorrect word or a [MASK] token, and then prompting the system to predict the correct word in its context by learning bidirectionally. Next sentence prediction helps BERT to learn the relationship between sentences, which, from a linguistic point of view, is important for the coherence of the language. Essentially, a binary decision is made about whether sentence B follows sentence A, which is done with the IsNext or NotNext tags.

The input of the token embeddings in BERT are multifaceted. On the one hand, WordPiece embeddings are used for token embeddings. Then, segment and sentence embeddings are stacked on top. Segment embeddings indicate whether a token belongs to sentence A or sentence B, and sentence embeddings indicate the position of the token in a sentence. All these embeddings have the same size and are trained simultaneously as opposed to previous neural approaches.

The embeddings are then passed to the encoder. First, each token embedding is passed to a fully-connected layer output, which has N number of neurons for N number of tokens. Then a softmax activation is applied to convert the word vector to a one-hot encoded vector representation for the word. We compare the one-hot encoded representation and the word vector, and train the network with the cross-entropy loss of the masked tokens.

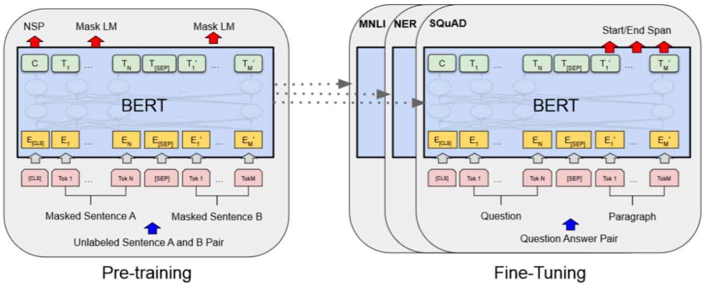


Figure 8: BERT Architecture (Devlin et al., 2019)

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a transformer with a similar architecture to BERT with a significantly increased amount of training data, 160GB, and improved computing efficiency (Liu et al., 2019). As opposed to BERT, it only relies on MLM for the pretraining and the loss for NSP is not considered in the training. It performs it in a dynamic way, so that the masking changes after each epoch while training the model, reporting better results for SQUAD 2.0 and SST-2 than BERT.

For our experiments, the BERT base[14] and RoBERTa base[15] language models were used via the Huggingface API. For this to work, the version 2.7.0 of the transformers

[14] https://huggingface.co/bert-base-uncased

[15] https://huggingface.co/roberta-base

library was installed. In short, the fine-tuning of these models corresponds to the training of a sequence labeller and text classifier for opinion target extraction and aspect category detection, respectively.

3.2 AllenNLP's Coreference Resolution

In 2020, Allen Institute for AI released a coreference resolution model[16] that relies on improved embeddings, the SpanBERT embeddings (Joshi et al., 2020), explained in section 2.4. These better account for context than previous approaches like GloVe embeddings. The SpanBERT system extracts coreferential spans, ranks them and prunes away the improbable candidates, using a higher-order coreference system based on Lee et al. (2018).

AllenNLP's coreference resolution takes as input a string of text, tokenises it with the SpaCy tokeniser, and returns an analysis of coreferential elements in the text. More specifically, the following elements are included in the output: antecedent_indices, clusters, document, predicted_antecedents, top_spans. The output for the predict function can be broken down in the following way:

- **antecedent_indices**: A list of possible antecedents for each coreference candidate. All coreference cluster candidates are possible antecedents for each candidate, hence the number of them is N^2.

- **clusters**: A list of coreference clusters, which each include token indices of the elements. Each element is marked with the start and end token of a coreference span, which are elements in the tokenised text in the document list.

- **document**: A list of the tokenised text. SpaCy is used by default for tokenising.

- **predicted_antecedents**: A list of predicted antecedents for candidates based on scores. The numbers refer to the index of each element in the top_spans list.

- **top_spans**: A list of candidates for coreference clusters.

In case we want to obtain a text with replaced coreferential mentions, this can be done with the replace_corefs function, which takes as input a list of coreference clusters, either the default list from the predictions or any custom list made up of token indices that we may want to pass to the function. Thus, when resolving coreference in a text, the indices of tokens have heightened importance, as all work is based on them. For our experiments, the version 2.1.0 of the AllenNLP library was used. A visual demo for AllenNLP's coreference resolutions model is available[17].

[16] https://docs.allennlp.org/models/main/models/coref/predictors/coref/

[17] https://demo.allennlp.org/coreference-resolution

3.3 Datasets

The dataset used in this thesis is the SemEval-2016 Task 5 on Aspect-Based Sentiment Analysis (Pontiki et al., 2016), which includes annotated reviews in eight different languages: English, French, Arabic, Dutch, Chinese, Russian, Spanish and Turkish, and for seven different domains: laptops, digital cameras, restaurants, hotels, museums, telecommunications, and mobile phones.

The English restaurant reviews[18] were chosen for this thesis, although French and Spanish were contemplated as well. First, coreference is more present in English than in Spanish, which tends to omit the subject instead. An example of this an be found in *The waiter was rude. It also took him ages to get our order.*, which translates to *El camarero era maleducado. Además, tardó muchísimo en tomar nuestro pedido.* and to *Le serveur était impoli. Et il a mis un temps fou à prendre notre commande.* We observe that the coreference in the second sentence is, in fact, omitted in Spanish, which leads to less coreferential elements. Also, the pronouns reveal the grammatical gender of the word, making the references slightly less ambiguous in Spanish (compare *lo, los* and *la, las* to *it* or *they/them*). French maintains the subject and retains more ambiguity in some of the pronouns (*l', les*) which could make it interesting for future work. In short, English was deemed interesting from a coreferential point of view due to the presence and ambiguity of the pronouns.

Category	Train	Test
AMBIENCE#GENERAL	255	66
DRINKS#PRICES	20	4
DRINKS#QUALITY	47	22
DRINKS#STYLE_OPTIONS	137	55
FOOD#PRICES	90	23
FOOD#QUALITY	849	313
FOOD#STYLE_OPTIONS	32	12
LOCATION#GENERAL	28	13
RESTAURANT#GENERAL	422	142
RESTAURANT#MISCELLANEOUS	98	33
RESTAURANT#PRICES	80	21
SERVICE#GENERAL	449	155
total	2507	859

Table 1: Aspect categories in the SemEval ABSA 2016 data set

The train set consists of 350 reviews, which have been annotated on a sentence-level. Altogether, there are 2000 sentences, out of which 1708 have annotated targets, category aspects, and polarities, and 292 have no annotations. Altogether 2507 targets were annotated, which means that some sentences have multiple targets. On average, there are 6

[18] https://github.com/howardhsu/ABSA_preprocessing/tree/master/dataset/SemEval/16/rest

sentences and 7 targets per review in the train set. On the other hand, the test set has 90 reviews with 676 sentences, out of which 587 have annotated targets, category aspects and polarities, and 89 sentences have no annotations. This means that, on average, there are 8 sentences and 10 targets per review.

In cases where it is not possible to locate the target of the opinion, it is considered implicit and the NULL tag is used for the annotation instead. However, an aspect category is always signed to a sentence if an opinion is present, which means we obtain a sentence-level classification for implicit targets as well. An in-depth analysis of the implicit targets is given in section 4.1.

All targets, both explicit and implicit, are annotated and classified into an aspect category, where the category (e.g. food) is given an aspect (e.g. prices), which takes on the form FOOD#PRICES. The aspect categories in the ABSA 2016 data set for English restaurant reviews can be seen in Table 2.

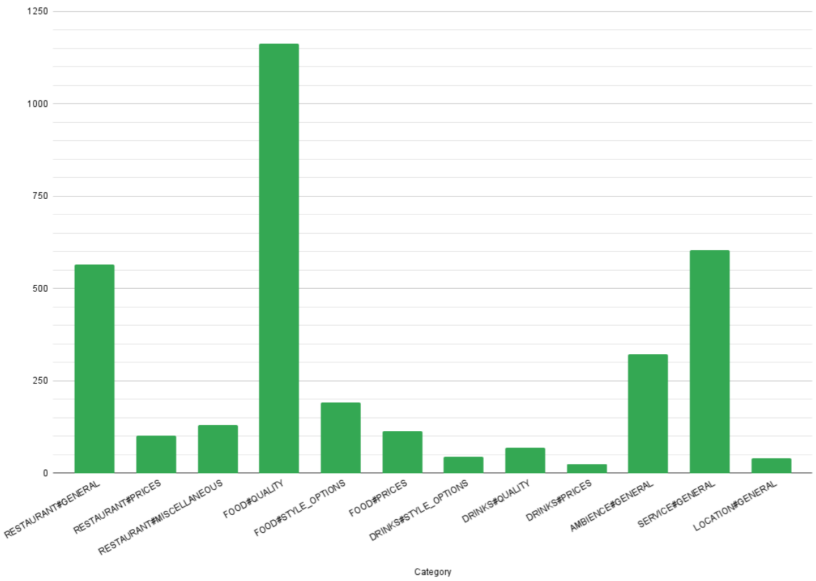


Figure 9: Number of targets in each category in the ABSA SemEval-2016 data set

3.4 Evaluation metrics

The main evaluation metric used for our experiments is the F1 micro-average score. It is computed as follows:

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In this case, F1 micro-average is preferred over F1 macro-average, because it calculates the F1 score for all classes together in the same set instead of calculating the F1 score per class, and then averaging the results. Using the F1 micro averaging results in higher scores in situations where class imbalance is present, as is the case in the ABSA SemEval-2016 data set (See Figure 9). This is because macro-averaging considers each class equally regardless of their sample number, whereas micro-averaging considers each sample. Micro-average is equal to accuracy, which can be formulated in the following way:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Additionally, the mean and standard deviation were calculated for experiments that were repeated, as the training batch varied in each epoch. This was done so as to balance out any differences in the results among data sets. The formulas are as follows:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \tag{1}$$

$$\mu = \frac{1}{n} \sum_{i=i}^{n} x_i$$

4 Experimental Setup

In this section, we explain in detail the experiments carried out with the material presented in Section 3. We analyse the targets in the data set, apply coreference resolution techniques on implicit targets, and perform opinion target extraction and aspect category detection. Our main goal is to improve opinion target extraction and aspect category detection by establishing corefential links and by using data with more explicit targets than in the original data set.

4.1 Implicit targets

An analysis of the targets was carried out in order to understand the scope of the implicit language in targets, annotated as "NULL" in the SemEval ABSA data sets. We observe that in the English restaurant review data set, the train set has 627 implicit targets and the test set, 209. These numbers account for approximately a quarter of the total number of targets, as seen in Figure 10. Thus, we consider it justified to apply different means of making explicit the implicit targets in order to maximise the information in the data set.

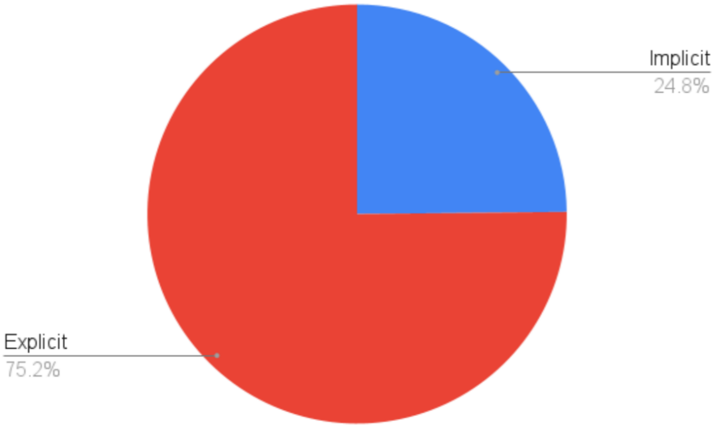


Figure 10: Explicit and Implicit Targets in the SemEval ABSA 2016 English Restaurant Reviews

To establish the differences in the nature of the implicit targets, a manual classification was deemed necessary. Each target was analysed in its sentence-level and document-level context and classified. If a suitable class did not exist, a new one was created during the analysis, which was conducted three times to ensure the targets were in correct classes.

This work was crucial for determining the number of coreferential targets that could be solved in our experiments. It may also explain the differences between the number of implicit targets in different categories. Figure 12 shows that a significantly greater number of implicit targets is found in the restaurant categories, with RESTAURANT#PRICES and RESTAURANT#GENERAL having more than half of the targets implicit. This suggests that customers tend to discuss these categories in a more abstract way using implicit language, whereas more restricted categories, like food and drinks, are discussed more explicitly.

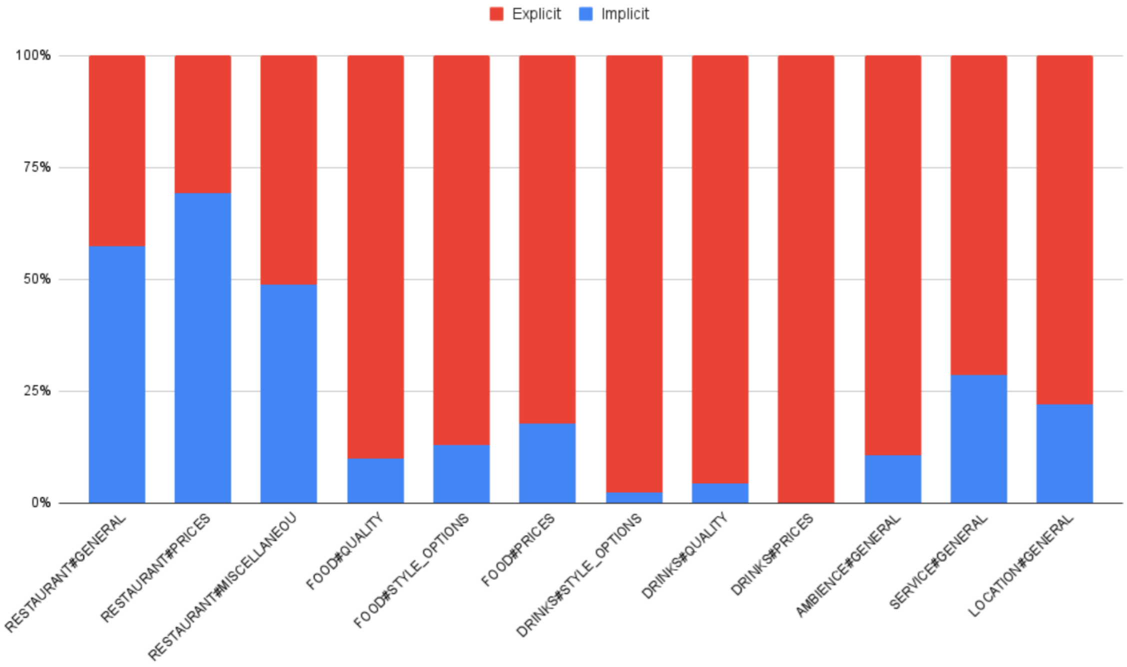


Figure 11: Distribution of Implicit Targets in Each Category for SemEval 2016 ABSA English Restaurant Data Set

Next, we outlined the criteria that was considered in the classification of implicit targets. It is worth noting that, although a one-to-one mapping was done between each target and class, there exists no direct correspondence between some targets and categories, as they are ambiguous and various classes could be correct.

- Coreference: The reference, usually a pronoun but also noun phrase in some cases, must have an annotated target within the document. Both anaphoric and cataphoric references were considered. These also include some locative references, such as *here* and *there*.

- Exophora: The reference is found outside the text and can be inferred by a human, but no coreference is possible.

- Unannotated coreference: A coreferential link exists, but the explicit reference is not a target.

- Omission: The target is omitted. There is no target information available.
- World knowledge: It is not possible to determine a target. The category and target are related to world knowledge, which are events, facts or other type of information, which requires common sense and reasoning.
- Guideline restrictions: Targets that could be extracted if the annotation guidelines[19] were different, such as the words *price* or *flavour*. Some verbs could be considered targets, like *serve*, but they are not entities, so they were ignored in the coreference analysis.

Type of target	Example	Explanation
Coreference	A weakness is the chicken in the salads. It's just average, just shredded, no seasoning on it.	The pronoun *it* refers to *chicken*.
Exophora	Ravioli was good...but I have to say that I found everything a bit overpriced.	*Everything* refers to every dish in the restaurant.
Unannotated coreference	My husband and I both ordered the Steak, medium. My husbands was perfect, my was well done and dry.	The explicit reference *steak* is not an annotated target.
Omission	A real dissapointment.	The writer does not mention what was a disappointment.
World knowledge	Their sake list was extensive, but we were looking for Purple Haze, which wasn't listed but made for us upon request!	The category and sentiment are based on an event.
Guideline restrictions	The flavors are amazing and the value is phenomenal.	Restrictions from annotation original SemEval guidelines. Otherwise *flavors* and *value* could be targets.

Table 2: Extracts of Each NULL Target Category

After using the classification methods outlined above and considering only a one-to-one mapping, the implicit target distribution is as visualised in Figure 12. We observe that coreferential targets account for over 20% of all implicit targets.

Coreferential references that do not have an annotated target make up around 5% of the data and could possibly be solved with coreferential methods, because in these experiments,

[19] https://alt.qcri.org/semeval2016/task5/data/uploads/absa2016_annotationguidelines.pdf

only references that had annotated targets were resolved. This means that if the implicit target's explicit target reference was in a sentence without a target or it was not annotated as a target, it was skipped. For example, *My husband and I both ordered the Steak, medium.* has no annotated targets, but is followed by *My husbands was perfect, my was well done and dry.* which includes two NULL targets referring to *the Steak*.

Exophoras are references outside the text, and are related to one of the following situations. Either the referred item exists outside the text and is ambiguous in the context, or the reference used is a general term which includes various items, such is the case of *everybody* or *everything*.

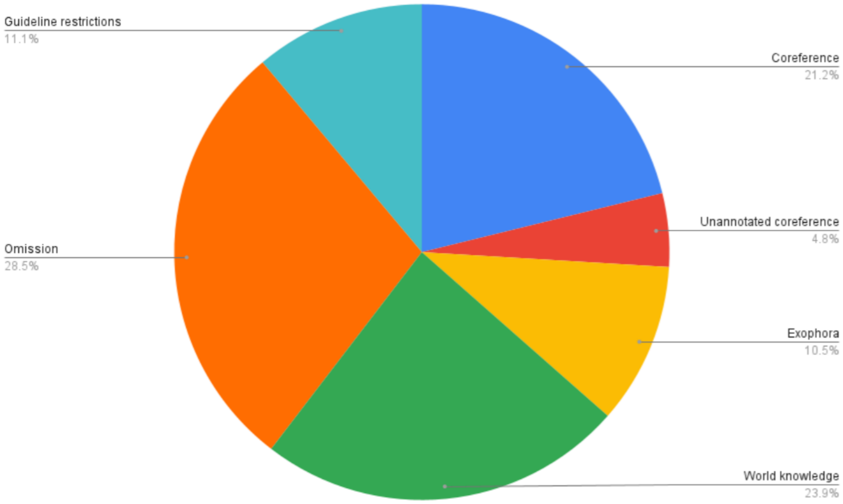


Figure 12: Distribution of Implicit Targets in SemEval 2016 ABSA English Restaurant Reviews

Regarding guideline restrictions, there is certain ambiguity concerning some of the NULL targets. Words like *experience* and *time* as well as possible verbal targets like *served* or *eat* have not been marked as targets, although they could be identified as such, as they could help to perform aspect category detection when the context taken into account. In the original annotations, the whole document was taken into consideration when annotating a sentence, which means that ambiguous words could be annotated and handled so that they could contribute to the improvement of opinion target extraction and aspect category detection.

4.2 Coreference resolution

After analysing the implicit targets and stating that approximately a fifth of them have coreferential links with targets, we outlined the coreference resolution experiments. Our main goal in this section was to link implicit targets with explicit targets and replace them in the reviews. To do so, two approaches were studied: automatic coreference resolution and manual coreference resolution.

4.2.1 Data preprocessing

In order to perform automatic coreference resolution, the data had to be processed and formatted first. We should consider that, although the original data is annotated on a sentence level, our experiments require the data to be organised on both document and sentence level. This is because, on the one hand, coreference resolution is performed on a document level, meaning that references are searched for in the entire document, and that, on the other hand, sentences containing NULL targets had to be identified. As the coreference resolution model we chose uses indexes to mark mention spans, we needed to obtain document-level indexes for the explicit ABSA targets and to identify all sentences with NULL targets so that coreference methods could be applied to the wanted data.

Considering the format of the original ABSA data set, the lxml library was used to extract the relevant information for the task (the texts, targets, and target character offsets). This step of information extraction was performed both on sentence and document-level. Then, in order to identify the explicit targets in the texts, we used the character offsets that are annotated in the original data set.

The texts were then tokenised with the SpaCy tokeniser using the en_core_web_sm pipeline. The SpaCy tokeniser was preferred for this task to avoid any mismatches in target indexes, as AllenNLP's coreference system also uses SpaCy tokeniser by default when it predicts the coreference mention spans. The indexes for all explicit targets were obtained on a document level, and they were needed for the manipulation of the output from the automatic coreference resolution.

4.2.2 Automatic coreference resolution

AllenNLP's coreference resolution model (See Section 3.1.2) was used to predict clusters of mentions referring to the same entity. The predict function[20] with the pretrained coref-spanbert model was used to establish links between implicit language and explicit language. When the coreference links are predicted, we obtain an output of clusters with mention spans, which indicate the start token and end token of each mention. A list of clusters is used in the replace function to determine the links, so our main focus was to create a list that includes valid clusters.

Not all automatic coreference clusters are correct and some are not relevant for this specific task. We wanted to establish links between annotated targets and coreferential NULL targets, which is why we maintained a cluster only if the following conditions were met:

- Cluster must have one target

[20] https://docs.allennlp.org/models/main/models/coref/predictors/coref/#predict

Element	Output
text	Great pizza for lunch place. Service was quick. The pizza was great. And it was quick which is very important. Have the iced tea. It was wonderful.
document	[‘Great’, ‘pizza’, ‘for’, ‘lunch’, ‘place’, ‘.’, ‘Service’, ‘was’, ‘quick’, ‘.’, ‘The’, ‘pizza’, ‘was’, ‘great’, ‘.’, ‘And’, ‘it’, ‘was’, ‘quick’, ‘which’, ‘is’, ‘very’, ‘important’, ‘.’, ‘Have’, ‘the’, ‘iced’, ‘tea’, ‘.’, ‘It’, ‘was’, ‘wonderful’, ‘.’]
clusters	[[[10, 11], [16, 16]], [[25, 27], [29, 29]]]
words in clusters	[[‘The’, ‘pizza’], ‘it’], [[‘the’, ‘iced’, ‘tea’], ‘It’]]
targets in review	pizza, service, NULL, iced tea
replaced text	Great pizza for lunch place. Service was quick. The pizza was great. And The pizza was quick which is very important. Have the iced tea. the iced tea was wonderful.

Table 3: An example of AllenNLP’s coreference resolution’s output

- Cluster must have sentences where NULL targets are present

After the coreference clusters were predicted, AllenNLP’s coreference replace function[21] was used to obtain texts where the references were replaced with explicit targets. The function takes as arguments a spacy document and a list of clusters. We only filtered out any unnecessary clusters to create a list of clusters that met the conditions outlined earlier in order to avoid any replacements of non-targets.

In addition, if several explicit targets were present in the cluster, only the first one was maintained. This was done in order to avoid the replacement of other existing explicit targets, as we only wanted to replace pronouns or other implicit language. Sometimes several lexical targets were present in a cluster due to their synonymous nature, such as in the case of *hostess* and *waitress*. Performing the steps above, the results for automatic coreference resolution are as reported by Tables 4 and 5.

Coreference status	Number of targets
Resolved	82
Unsolved	58
Total	140

Table 4: Resolved Targets in Train Set

4.2.3 Manual Coreference Resolution

Although AllenNLP’s coreference resolution model was able to establish correct links for 59% of the targets classified as coreferential in the train set and 25% of the targets in the

[21] `https://docs.allennlp.org/models/main/models/coref/predictors/coref/#replace_corefs`

Coreference status	Number of targets
Resolved	9
Unsolved	26
Total	35

Table 5: Resolved Targets in Test Set

test set, we wanted to resolve all of them in order to get a better understanding of the effect of applying coreference resolution to ABSA data for aspect category detection and opinion target extraction. Manual coreference resolution was then performed in order to compare the results of opinion target extraction and aspect category detection with different types of data: the original ABSA 2016 data set, automatic coreference resolution and manual coreference resolution.

As the line number for each NULL target was annotated in the classification process, the coreferential NULL targets could be easily reviewed manually by searching for the corresponding lines in the original xml document. Missing links were established by replacing noun phrases and pronouns with the explicit target. Similarly, any incorrect coreferential links were deleted or corrected. A link was considered incorrect if it was established where it was not applicable or, in more ambiguous cases, if the link was grammatically correct but its annotated category did not correspond to the target. The links were searched for in the whole document, although the references were usually found in the previous sentence.

Moreover, if any links required an additional word like a preposition in order to be grammatical in the sentence, it was added. For example, in the sentence *This place is worth an one-hour drive.*, the predicted target was *This place*, which means that using the target to replace the pronoun *here* in the following sentence *I am so coming back here again, as much as I can.* would give us *I am so coming back This place again, as much as I can.*. We decided to add any missing prepositions so as to make the sentences grammatical. In fact, this type of cases were not handled at all in the automatic coreference resolution and are analysed more in depth in Section 6 along with other problematic cases.

After performing manual coreference resolution, we were able to resolve all 140 implicit targets classified as coreferring targets. Doing so increased the number of explicit targets in the data set from 2530 to 2705 (Figure 13).

4.3 Opinion Target Extraction and Classification

Opinion target extraction (OTE) intends to identify all the targets in a text for further processing, such as classification into broader aspect categories or into sentiment categories. With supervised methods, OTE is usually performed as a sequence labelling task in order to capture multi-word targets. Apart from performing OTE, we also carry out an OTE experiment with sequence labelling that classifies the targets into aspect categories. Our experiments are carried out with the original ABSA 2016 data, with the automatically resolved coreference data and with the manually resolved coreference data, as the goal is to study the effect of coreference resolution and to understand whether the sequence

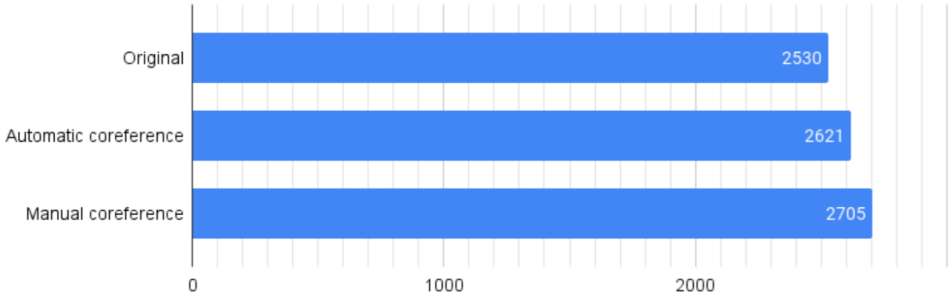


Figure 13: Explicit Targets in Each Data Set

labelling system improves by making explicit the implicit targets.

First, the data was formatted into a tab-separated values files so that each row contained a word and its respective BIO tag separated by a tabulation. The BIO tagging scheme (Table 6) was used to annotate the words and two types of files were generated: one annotating the opinion targets with the BIO scheme, and another annotating targets with the BIO scheme and aspect category. These files were generated for the original data, automatically resolved coreference data and manually resolved coreference data.

The	wine	list	is	interesting	and	has	many	good	values
O	B-TARGET	I-TARGET	O	O	O	O	O	O	O

Table 6: BIO Tagging Scheme

A significant part of the coreferential targets predicted by AllenNLP were noun phrases starting with a determiner. This means that also the determiners were tagged as part of the target, even though they are semantically empty words. We wanted to align the targets with the original targets, which is why we deleted the BIO tags from the determiners starting the targets. This was done by looking for target spans starting with any of the following words: a, an, the, his, her, their, that, this, these. Targets including any of these determiners in the middle of the span, such as *view of the new york city skiline* were left intact, as they are part of the target in the original annotations as well.

Sentence	Target	Target type
The place is a lot of fun.	place	original annotation
My six year old loved it. My six year old loved The place	NULL The place	original annotation AllenNLP prediction

Table 7: AllenNLP's Predicted Target with a Determiner

Sequence labelling was performed on both target annotated and aspect category annotated files with the same model and parameters. Transformer-based models RoBERTa base and BERT base uncased were used with a script[22], as detailed in Section 3. These models do not require feature engineering like previous machine learning based systems, which is why no features were pre-selected for the documents.

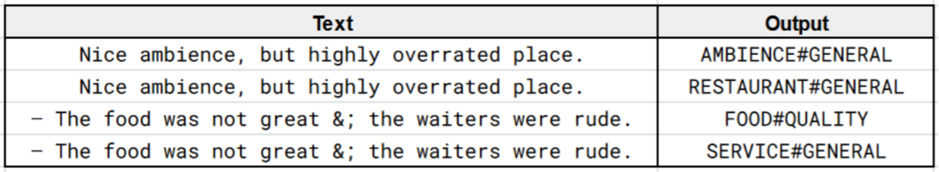
All sequence labelling experiments were carried out for 5 and 10 epochs in order to compare the performance of the systems. The model was trained with a cloud GPU and the hyperparameters were the following: maximum sequence length was set to 128, the batch size was 32, and the learning rate was 5e-5, as recommended by Agerri and Rigau (2019). The seed was set in order to generate reproducible results.

4.4 Aspect Category Detection

Aspect category detection is often treated as a text classification task in which the opinion targets are classified into a category. We decided to apply two different approaches to ACD: multi-class text classification and a multi-label text classification. Multi-class classification means that any text that has several labels is multiplied by the number of labels and each text is assigned one label (Figure 14). However, multi-label classification approaches the problem from another point of view: the texts are left as they are and they are assigned various labels (Figure 15).

In order to perform multi-class classification, the data was formatted into tab-separated values files, where each text was preceded by a tab and a class. In cases where the text had various labels, the text was simply multiplied by the number of labels and assigned one label. Any sentences that were lacking opinion targets were deleted from the data set. Following the same evaluation method as the original SemEval evaluation guidelines (Pontiki et al., 2016), a set was created of all the texts, which means any identical texts with identical tags were removed. This type of cases are numerous for texts with the FOOD#QUALITY tag, as it often refers to many different dishes.

For multi-class classification, the transformer-based RoBERTa and BERT models were used with a script [23]. Their specifications can be checked in Section 3. As for sequence labelling, the maximum sequence length was 128, the learning rate was 5e-5 and the batch size was 32. The classification was performed for 5 and 10 epochs, and the seed was set, in order to compare the performance of the systems.

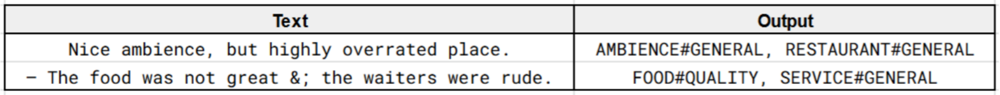


Text	Output
Nice ambience, but highly overrated place.	AMBIENCE#GENERAL
Nice ambience, but highly overrated place.	RESTAURANT#GENERAL
- The food was not great &; the waiters were rude.	FOOD#QUALITY
- The food was not great &; the waiters were rude.	SERVICE#GENERAL

Figure 14: Example of Multi-Class Classification

[22] https://github.com/ragerri/transformers-training-scripts/blob/master/run_conll_ner.py

[23] https://github.com/ragerri/transformers-training-scripts/blob/master/run_classification.py



Text	Output
Nice ambience, but highly overrated place.	AMBIENCE#GENERAL, RESTAURANT#GENERAL
- The food was not great &; the waiters were rude.	FOOD#QUALITY, SERVICE#GENERAL

Figure 15: Example of Multi-Label Classification

Multi-label experiments were carried out with an adapted version of an existing notebook[24]. As multi-label classification outputs various tags for one text, it is best to use a binary matrix indicating the presence or non-presence of a class for each text. Therefore, the data was formatted into a matrix with binarised category labels using the Multi-label binarizer from Scikit learn [25].

The training was done on both RoBERTa and BERT with a batch size of 32 and a maximum length of 100. All experiments were carried out for 5 and 10 epochs like the previous experiments. However, as the seed was not set in this script, the batches varied, so the training with each different model, data set and epoch number was repeated five times. Thus, the final results reported for every experiment are the average of the 5 runs.

[24] https://colab.research.google.com/github/rap12391/transformers_multilabel_toxic/blob/master/toxic_multilabel.ipynb

[25] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html

5 Empirical Results

This section shows the results for the experiments outlined in the previous section for Opinion Target Extraction and Aspect Category Detection. We explain the quantitative results for each model that we trained and draw preliminary conclusions from them.

5.1 Opinion Target Extraction and Classification

As explained in Section 4.3, we trained models for two types of sequence labelling models: one for opinion target extraction and another one for opinion target extraction with aspect category classification. Both type of models were trained with three data sets: the original ABSA data, the automatically generated coreference resolved data and the manually resolved coreference data.

The results for opinion target extraction show that, overall, coreference resolution is beneficial for sequence labelling. As observed in Table 8, the best results are obtained by training with RoBERTa base for 10 epochs with a manually resolved coreference data set, which gives us an F1-score of 0.8168. This is nearly three points more than the results with the original ABSA 2016 data set. We also observe that the results improve when using BERT regardless of the coreference type. For RoBERTa, only manual coreference improved the performance, as automatic coreference resolution presumably introduced noise in the training, which worsened the F1-score compared to the models trained with the original data. The greatest improvement is found in the model trained with BERT for 5 epochs, as we observe an improvement of nearly 8 points from the original data set to the manually resolved coreference data set.

System	Original data	Automatic coreference	Manual coreference
BERT base uncased (5)	0.6886	0.728	**0.7635**
BERT base uncased (10)	0.6866	0.738	**0.7528**
RoBERTa base (5)	0.7811	0.7435	**0.7815**
RoBERTa base (10)	0.781	0.7494	**0.8168**

Table 8: Results for Opinion Target Extraction

Coreference resolution has less overall improvement when opinion targets are extracted and classified into aspect categories via sequence labelling (See: Table 9). The best results were obtained with the manually resolved coreference data set (0.6697) but they are only slightly better than the best results for the original ABSA 2016 data set with RoBERTa. We observe that, generally, automatic coreference resolution is helpful when used with BERT, but not when used with RoBERTa. Manual coreference, however, helps obtain better performance, with both BERT and RoBERTa, except for RoBERTa that was trained for 10 epochs. The greatest difference in the results can be found in the model that was trained with BERT for 5 epochs, as we only obtain an F1-score of 0.5571 with the original data, and 0.6428 with the manual coreference data, an improvement of approximately 9 points.

System	Original data	Automatic coreference	Manual coreference
BERT base uncased (5)	0.5571	0.596	**0.6428**
BERT base uncased (10)	0.5817	0.6435	**0.6697**
RoBERTa base (5)	0.5974	0.5701	**0.61**
RoBERTa base (10)	**0.6634**	0.5549	0.6424

Table 9: Opinion Target Extraction and Aspect Category Detection as Sequence Labelling

5.2 Text Classification for ACD

Aspect category detection was handled with two different experiments: multi-class classification and multi-label classification, as detailed in section 4.4. The results for both are presented separately, after which a comparison is added to evaluate both methods used.

The results for multi-class classification do not differ drastically among the data sets and systems. The best results were obtained with BERT trained for 10 epochs with the automatically resolved coreference data set, which gave us an F1-score of 0.6927, and the RoBERTa model which was trained for 10 epochs with the manually resolved coreference file, which also obtained an F1-score of 0.6927 (See Table 10). The greatest improvement is observed in BERT trained for 10 epochs, where the results improved approximately 4 points from the original data set to the automatically resolved coreference data set. We could conclude that coreference resolution is not that relevant for multi-class classification.

System	Original data	Automatic coreference	Manual coreference
BERT base uncased (5)	**0.6712**	0.6617	**0.6712**
BERT base uncased (10)	0.655	**0.6927**	0.69
RoBERTa base (5)	**0.69**	0.6617	**0.69**
RoBERTa base (10)	0.6873	0.6792	**0.6927**

Table 10: F1 scores for our Multi-class Classification Models

System	Original data	Automatic coreference	Manual coreference
BERT base uncased (5)	0.7354±.0270	0.7606±.0345	**0.7628±.0306**
BERT base uncased (10)	0.7777±.0147	0.7745±.0086	**0.7913±.0119**
RoBERTa base (5)	0.778±.0328	0.7973±.2977	**0.802±.0340**
RoBERTa base (10)	0.8192±.0120	0.8094±.0082	**0.8246±.0157**

Table 11: F1 Score (Mean and Standard Deviation) for Our Multi-Label Classification Models

The results for multi-label classification suggest that coreference resolution improves aspect category detection. As opposed to multi-class classification, the training of the models for multi-label models was repeated five times for each system, as the seed was not set and the batch varied. Thus, the average of the all runs was calculated for the evaluation. As we can observe in Table 11, manually resolved coreference data helps to classify texts and perform aspect category detection. The best results are obtained with RoBERTa trained for 10 epochs, with which we obtain an F1-score of 0.8246. This is a great improvement compared to the previous multi-class experiment, and we can conclude that manual coreference resolution helps, especially when less epochs are used to train the model. We observe that the use of automatic and manual coreference resolution has a similar effect for multi-class classification and multi-label classification (Figure 16).

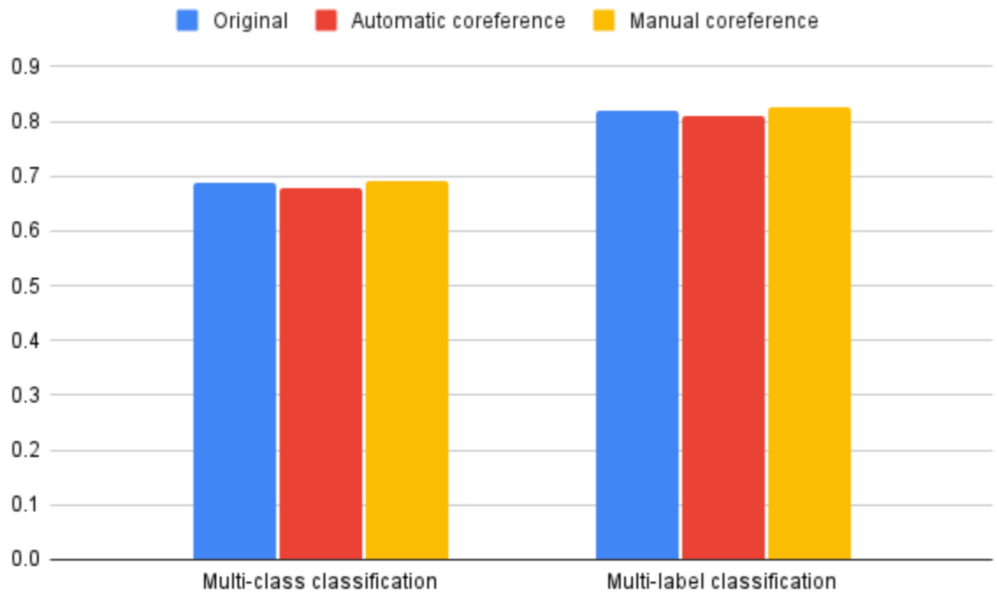


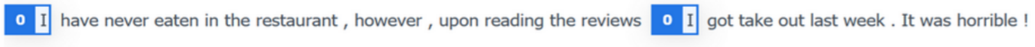
Figure 16: F1 scores for RoBERTa Trained for 10 epochs for Aspect Category Detection

6 Discussion and Error analysis

This section aims to give a more in-depth analysis of whether and how coreference resolution improves the systems used for OTE and ACD. The discussion is structured as follows: linguistic theory of coreference is presented together with coreference errors, then an analysis of the errors in our opinion target extraction model is given, followed by an analysis of the differences found in our aspect category detection model.

6.1 Problematic Cases in Coreference

Although this thesis does not aim to improve coreference systems, it is still important to understand the type of errors and their origin in the data set used for OTE and ACD. This way, we can understand the limits of automatic coreference resolution and evaluate whether automatic or manual coreference resolution is useful for the ABSA tasks. After analysing errors found in the AllenNLP coreference system, we discuss the performance of our sequence labelling and text classification models in detail.



0 I have never eaten in the restaurant , however , upon reading the reviews 0 I got take out last week . It was horrible !

Figure 17: Ambiguous “it” in AllenNLP’s Coreference System

Cataphora. According to Merriam-Webster, a cataphora is “the use of a grammatical substitute (such as a pronoun) that has the same reference as a following word or phrase”. In the case of cataphoras in automatic coreference resolution, the lexical replacement is incorrect, as a pronoun is used to replace explicit targets. AllenNLP’s coreference resolution system places the first mention of the references, whether it is a pronoun or explicit item, as the head of the cluster. New functions can be written to overcome this issue, but it is worth noting that AllenNLP does not have a simple way - such as a separate function or a parameter - to look for cataphoras or avoid the usage of pronouns as head of cluster. We find examples like *Loved it.* followed by *I must say I am surprised by the bad reviews of the restaurant earlier in the year, though.*, where the correct reference cluster is found, but *the restaurant* is erroneously replaced by *it*. This, in fact, results to be counterproductive, as the pronoun is then used for lexical replacement makes implicit a target.

Deixis. Deixis is a linguistic concept that explains the variable nature of place, person and time. Although time may not be relevant for OTE, persons and places are often referred to in reviews. Deictic expressions such as *here* or *there* always depend on the speaker’s point of view in time and cannot be resolved with coreference resolution if there is no explicit reference in the review. This is extremely complicated for automatic systems to solve even if an explicit referring target is present, as none of the deictic references of place were solved by AllenNLP’s coreference model. This could be due to the fact that a preposition is needed in order to introduce the place explicitly. These cases were resolved manually, as discussed in Section 4, which means that *I would defiantly come back here again as one of my top choices.* became *I would defiantly come back to the place again as one of my top choices.*, as *place* is an annotated target in the review.

Synecdoches. Targets that have a synecdochic reference are rarely correctly resolved by the model. For example, *The place is the next best thing to my Moms cooking.* has one NULL target with the category FOOD#QUALITY. However, there is no explicit mention of food, so we assume the target is *The place.* We can consider that the writer refers to the restaurant's food and not to the establishment itself. These cases are extremely difficult to resolve with automatic coreference systems, but also with manual coreference. For the gold annotations, *The place* was replaced with the antecedent target *food* in order to match the category more explicitly. This helped the model to predict the correct category in comparison to the original data or automatically resolved data.

Distance from reference. From a cognitive perspective, a hearer is more likely to connect closer references together than ones with a long distance, but can make sense of the text with common sense. For the automatic systems, long distances between coreference cluster elements are a great problem. For example, the target *chicken* in the sentence *I took one look at the chicken and I was appalled.* was not correctly linked to *So I decide to report back to the waitress because it was completely inedible.*, as other dishes were discussed between the sentences.

Ambiguity. Ambiguous pronouns can be problematic to process for humans, but for machines the task of handling ambiguity is even more complicated. This lead to incorrect clusters when several candidate references were present. For example, in *I highly recommend it.*, the pronoun *it* was replaced by *food*, although the annotated category was RESTAURANT#GENERAL and the real reference was *Jekyll and Hyde.* Figure 17 presents such an ambiguous case that even a human could link *it* to either the take away or the whole experience, or interpret it as an expletive *it* with no semantic value.

Several coreference clusters. Errors were found due to overlapping coreference clusters. Often from a syntactic point of view, the referring pronoun could have several candidate coreference links, but from a discourse point of view, humans are able to distinguish what is being referred to. For example, in *We were not dissappointed in the least bit by this little gem* the predicted target *this little gem* was linked to *The bagel* in *The bagel was huge* and to *They* in *They were served warm and had a soft fluffy interior.* We observe that the number does not even correspond in the reference *they.*

Similar references. Related to overlapping coreference clusters, any entities that are similar are easily confused with the implicit pronoun. For example, both the dishes *salmon* or *chicken* could be referents for ”it”, and both *hostess* and *waitress* could be referents for ”she”. If semantically similar words appear close to each other in a text, there is a chance that the machine chooses the wrong referent, although the reference is clear for a human.

It is important to note that not all coreference links are useful for aspect category detection even if they contain targets and a correct link. For cases like *The food was all good but it was way too mild.* followed by *Normally, places ask how hot you want it, but they didn't.* a coreferential link can be established between *the food* and *it.* However, since the second sentence's NULL target is annotated as service category and not food category, solving a link between *The food* and *it* would not improve the system.

6.2 Analysis of Opinion Target Extraction

A comparison of sequence labelling with original data, automatically resolved coreference data and manually resolved coreference data was carried out with both quantitative and qualitative measures. For these analyses, the output of the best system was chosen. This means that the analysis for simple sequence labelling is based on the RoBERTa base 10 epochs model and the analysis for sequence labelling with a category is based on BERT base uncased 10 epochs models.

6.2.1 Sequence Labelling for OTE

Both quantitative and qualitative differences were found for the three models. First, Figure 18 shows that there is a correlation between the use of a coreference model (explained in Section 4.2) and between the number of correctly predicted opinion target expressions. Only the output for coreferential targets was analysed in the predictions, so any changes outside of these targets were not taken into consideration. A label was considered only partially correct if it was a multi-word target and some of the words tagged correctly or if the label started erroneously with an I-tag instead of B-tag. In total, 25 out of the 35 coreferential targets were solved with the manually corrected coreference data, 13 with the automatic coreference data, and only three with the original data.

Frequent targets. One of the greatest errors we encountered was the overtagging of common words like *food*, *place* and *restaurant* when these were not real targets of the sentences. In some cases, these aforementioned frequent words were not tagged as targets, although they should have been. Our coreference models seemed to improve this aspect. For example, the model with the original data would tag *This* as a target in the sentence *This is a great place to get a delicious meal!*, whereas both of our coreference systems would tag *place* as a target. Considering this, it would be advisable to train a model with data that has varied opinion targets to avoid overtagging frequent words.

Predictions		Gold	
saag	B-TARGET	saag	B-TARGET
and	I-TARGET	and	O
paneer	I-TARGET	paneer	B-TARGET
and	I-TARGET	and	O
korma	I-TARGET	korma	B-TARGET

Table 12: Example of Errors in Target Sequence Lengths

Opinion target length. Another issue for our sequence labelling models was distinguishing various short targets from a long target. Table 13 shows an example of this type of tagging error, where three targets are grouped as one target. Although this may seem like a triviality, it would be preferable to extract each target separately each corresponding to a different dish, so that they can be further processed.

Cataphoras. Next, we observe the effect of not fixing the lexical replacement in cataphoras. The pronoun *they* is erroneously tagged as a target some contexts where it is used as a generic pronoun. Although not many instances of target-tagged pronouns were found in the automatic coreference predictions, having pronouns tagged as targets could introduce noise when training the model.

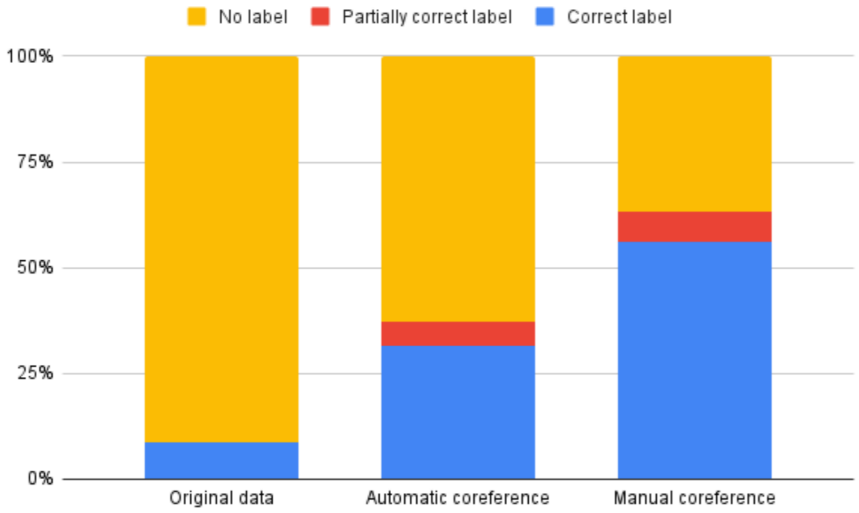


Figure 18: Target Predictions for Coreferential Targets

Proper names. In addition, some proper names were tagged correctly in the models trained with resolved coreference data, as opposed to the original data that did not handle them correctly. For example, Mioposto, name of a restaurant, was only tagged as a target in our coreference resolved models as opposed to the original data. On the other hand, McDonald's was erroneously tagged as a target with the original data, but not with our data sets with resolved coreference links. Lastly, some functional words like *not* were tagged as a target in the original data, but not with the coreference data.

Lastly, we observe general consistent improvements in the systems trained with coreferential data. Every time a lexical replacement of a target was fixed, this was usually tagged. For example, the original data had the word *They* referring to onion rings untagged, whereas our coreference data had the replacement *onion rings*, which was tagged as a target. Similarly, if the original data contained the word *she* and this was replaced by *the waitress*, *waitress* was tagged as a target.

6.2.2 Sequence labelling with OTE and Aspect Category

Extracting opinion targets with their category is a significantly more complex task than simply extracting the opinion target. Errors found in this task include all the previous errors, but also incorrect categories. First a graph for coreferential target predictions for each system is analysed and then a qualitative analysis of the errors and improvements is presented.

In Figure 19, we observe that using a data set with resolved coreference links helps to extract opinions. For tags containing the aspect category, a label was considered partially correct if the entity (food, restaurant, service...) was correct and the attribute (general, prices, miscellaneous...) was incorrect. As with the simple target labels, a target was considered only partially correct if if started with an I-tag instead of B-tag. Thus, only targets containing the whole aspect category correct (i.e. entity and attribute) were considered correct. The model trained with the original data reports only 1 correctly predicted coreferential target, the automatic coreference has 9 correctly predicted coreferential targets and the manually resolved coreference has 17 correctly predicted targets. The results are, thus, similar to the simple opinion target extraction in Section 6.2.1.

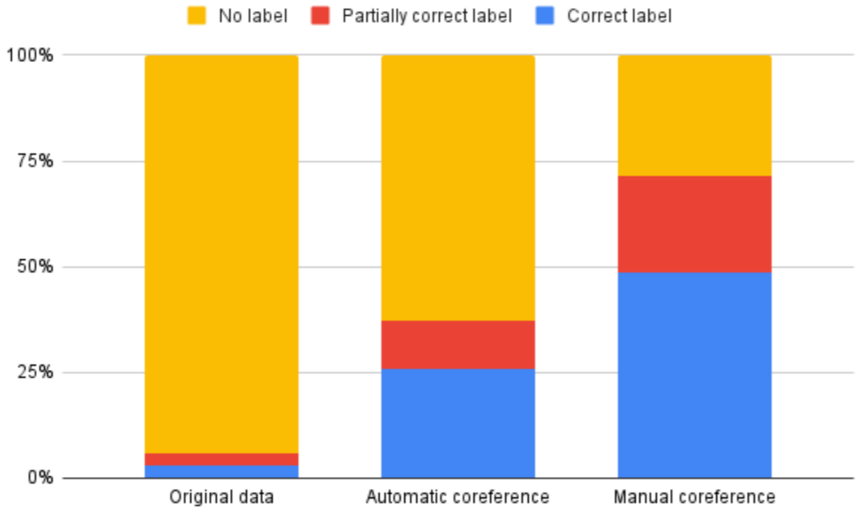


Figure 19: Opinion Target Predictions and ACP for Coreferential Targets

Mismatch of attributes. An interesting case is the tagging of *onion rings*, where *onion* is tagged with B-FOOD#QUALITY but *rings* is tagged with I-FOOD#STYLE_OPTIONS. From a semantic point of view, this makes sense when analysing the words separately, but we should keep the same tag for both words, as *onion rings* constitutes a different dish from onions.

Synecdoches. Replacing synecdochic references with words that fit into the category helped us perform sequence labelling with correct categories. This was the case for *Mercedes* *restaurant is so tasty, the service is undeniably awesome!* where Mercedes restaurant was replaced by another target from the same review, *catering*.

6.3 Analysis of Aspect Category Detection

Both multi-class and multi-label experiments were analysed with quantitative and qualitative methods. Only the test predictions for the best results were analysed, which means RoBERTa base 10 epochs was chosen for both the multi-class classification and the multi-label classification analysis.

6.3.1 Multi-class Classification

Table 13 shows differences for each category and data set. We observe that the original data performs better for RESTAURANT#GENERAL, RESTAURANT#MISCELLANEOUS and, slightly better for AMBIENCE#GENERAL as well. It is interesting to note that these correspond to some of the categories that had the highest number of implicit targets in them, as mentioned in Section 4.1.

When comparing the results, the automatic coreference does not highlight the performance for any specific categories; two of the categories that performed the best are shared with the manually resolved coreference results and DRINKS#PRICES only has three instances in the test set (detailed in Section 3.1.), which is why that particular category is not of great interest. The results for manually resolved coreference, however, suggest that the classification for many of the categories improved, especially for the greatest implicit category FOOD#PRICES. The results seem to be directly proportional to the number of samples in each category, with the most common categories gaining better results. The largest categories (FOOD#QUALITY, RESTAURANT#GENERAL and SERVICE#GENERAL) obtain the best results.

We analysed the predictions of all targets that we classified as coreferential in the test set. Figure 20 shows that the original data set has 25 out of the 35 coreferential target texts correctly classified, the automatic coreference has 27 correctly classified texts and the manual coreference has 26 correctly classified texts. The results are, thus, very similar and we can conclude that automatic coreference resolution does not help to perform aspect-category detection with multi-class classification methods. Manual coreference resolution improves the system very slightly, from 0.6873 to 0.6927. However, in order to understand which factors influence the classification of texts with implicit coreferential targets, we analysed each case separately. The following conclusions could be drawn:

Multi-target sentences. One of the greatest problems was tagging sentences containing various opinions (Table 14). Coreference resolution had no effect on aspect category detection when several opinions were present in a text. Moreover, multi-target sentences were assigned only one tag also when all the targets were explicit.

Locative adverbs. Although for OTE replacing locative adverbs, such as *here* and *there*, helped to extract opinion targets, this replacement seemed irrelevant for ACD. Texts

Category	Original	Automatic	Manual
AMBIENCE#GENERAL	**0.6569**	0.6250	0.6560
DRINKS#PRICES	0.0000	**0.5714**	0.4000
DRINKS#QUALITY	0.4000	**0.5000**	**0.5000**
DRINKS#STYLE_OPTIONS	0.1538	**0.3529**	**0.3529**
FOOD#PRICES	0.3830	0.3137	**0.4151**
FOOD#QUALITY	0.7580	0.7442	**0.7634**
FOOD#STYLE_OPTIONS	0.4103	0.4103	**0.4889**
LOCATION#GENERAL	0.4444	0.4211	**0.4545**
RESTAURANT#GENERAL	**0.8519**	0.8390	0.8370
RESTAURANT#MISCELLANEOUS	**0.3137**	0.2800	0.2759
RESTAURANT#PRICES	0.4242	0.4848	**0.5263**
SERVICE#GENERAL	0.7391	0.7412	**0.7516**
Micro-average	0.6873	0.6792	**0.6927**

Table 13: F1 Score for Each Category in Our Best Model for Multi-Class Classification

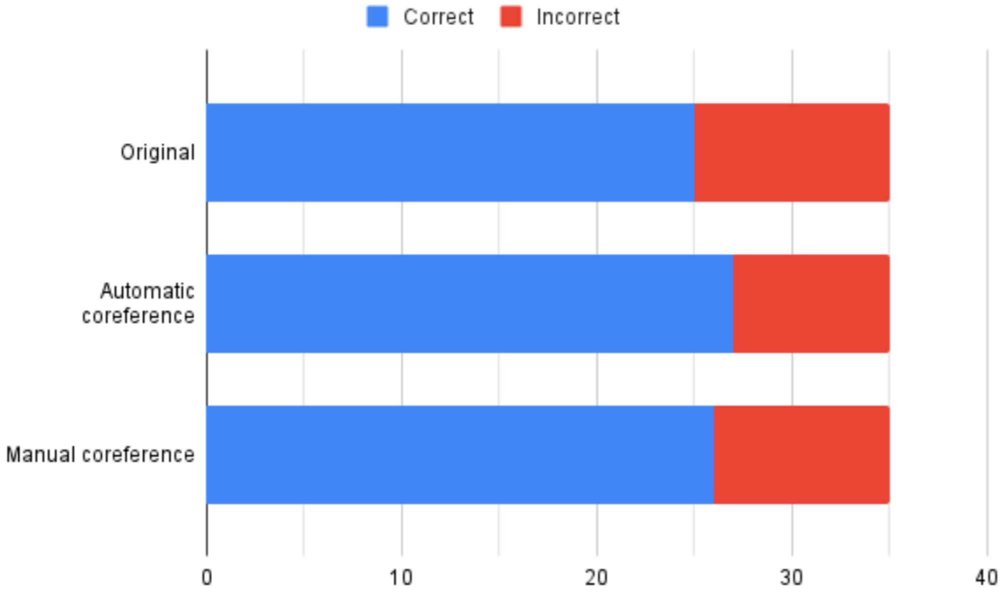


Figure 20: Number of Classified Sentences with Coreferential Targets with Multi-Class Methods

with implicit places were classified correctly, as were texts with explicit places, such as *restaurant* or *place*.

Personal pronouns. As with locative adverbs, it seems that replacing words like *she* or *they* does not have an effect on aspect-category detection, as these were correctly

"I will never forget the amazing meal, service, and ambiance I experience at this restaurant."			
Original	FOOD#QUALITY	FOOD#QUALITY	FOOD#QUALITY
Automatic	FOOD#QUALITY	FOOD#QUALITY	FOOD#QUALITY
Manual	FOOD#QUALITY	FOOD#QUALITY	FOOD#QUALITY
Baseline	FOOD#QUALITY	SERVICE#GENERAL	AMBIENCE#GENERAL

Table 14: Tagging Error for Multi-Target Sentences with Multi-Class Methods

classified also with the implicit targets.

Ambiguous *it*. It seems like most improvements - either with the automatic or manual coreference - come from the replacement of the pronoun *it*. Such is the case for *Don't leave the restaurant without it.* or *BUt once done, it's not too much dough, not too much cheese, not too much sauce.* which were incorrectly classified with the implicit target, but correctly classified with explicit target replacements. In both cases, these are improvements in the FOOD#QUALITY category.

Unclear aspect. Some errors stem from the subtle difference between the categories RESTAURANT#GENERAL and RESTAURANT#MISCELLANEOUS. This is also reflected in the texts containing coreferential targets.

6.3.2 Multi-label Classification

As mentioned before, the best performing model RoBERTa base trained for 10 epochs was chosen for the analysis. We observe that using coreference resolution with multi-label classification does not increase the F1-score significantly, as opposed to multi-class classification, if we analyse the results for each category (Table 15). Multi-label classification scripts were run five times for each data set, so we randomly chose one output of each data set for this analysis.

It must be highlighted that the difference in F1 score for the original, automatically resolved coreference data and manually resolved coreference data are not very noticeable in the classification results (Table 15). However, the overall results are significantly better than for multi-class classification. There is some variance among the performance for different categories but no correlation between coreference resolution is found.

To better understand the errors and differences, we calculated the number of coreferential targets that were correctly classified (Figure 21). We observe that among the coreferential targets, the model trained with manual coreference resolution and the original data perform better than the model trained with automatic coreference resolution, as the first two have 31 correctly classified targets out of 35, and the latter has only 27. As with the multi-class classification results, we analysed the predictions of all targets that were classified as coreferential in the test set.

Ambiguous it. The greatest difference we observed is related to the pronoun *it*. We find that some of these "its" were classified into several classes with the original data, although they only belonged to one. For example, *Don't leave the restaurant without it* was correctly classified into FOOD#QUALITY with all data sets, but the original data

Category	Original	Automatic	Manual
AMBIENCE#GENERAL	**0.84**	0.81	0.82
DRINKS#PRICES	0.00	0.00	0.00
DRINKS#QUALITY	0.33	**0.67**	0.56
DRINKS#STYLE_OPTIONS	0.00	**0.35**	0.25
FOOD#PRICES	**0.72**	0.68	**0.72**
FOOD#QUALITY	0.91	0.91	**0.92**
FOOD#STYLE_OPTIONS	**0.59**	0.54	0.57
LOCATION#GENERAL	0.00	0.44	0.00
RESTAURANT#GENERAL	0.82	**0.83**	0.80
RESTAURANT#MISCELLANEOUS	0.43	0.34	**0.45**
RESTAURANT#PRICES	0.52	0.63	**0.68**
SERVICE#GENERAL	**0.91**	0.89	0.90
Micro-average	0.80	0.80	0.80

Table 15: F1 Score for Each Category in Our Best Model for Multi-label Classification

also classified it into RESTAURANT#GENERAL erroneously. This is also the case for the sentence *It was absolutely amazing.*, as both coreference resolved models managed to assign it only one category, RESTAURANT#GENERAL, whereas the original data classified it also in FOOD#QUALITY.

Personal pronouns and locatives. Replacing implicit targets referring to people, such as *she* or *they*, or locative adverbs, such as *here* or *there*, does not help to classify texts. This was also the case for our multi-class classification results.

Multiple categories. One of the main drawbacks of multi-class classification seemed to be the assignation of only one category where multiple were present (See Table 15). Multi-label classification seems to address this problem, and classifies the example of Table 15 into three categories. However, sometimes texts are over-assigned more categories than there should be, although these do include the correct category as well.

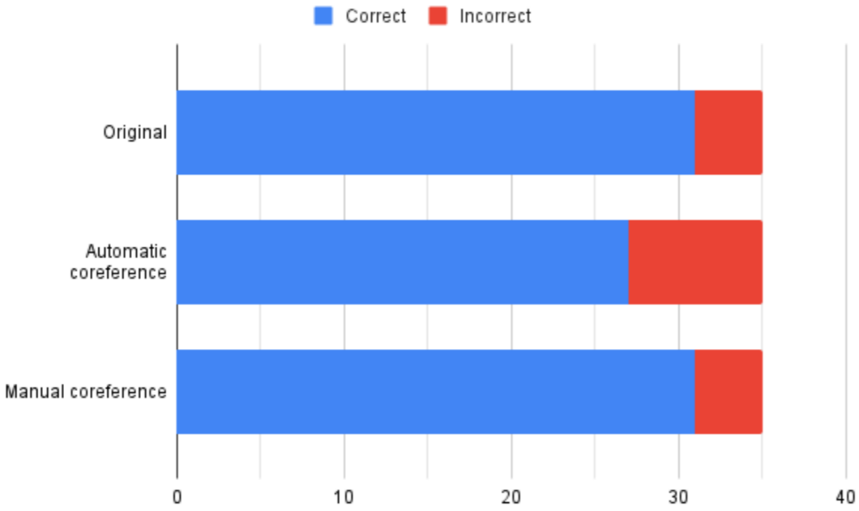


Figure 21: Classified Sentences with Coreferential Targets with Multi-label Methods

7 Conclusion

This thesis has explored techniques to handle implicit language in two of the main ABSA tasks: opinion target extraction and aspect category detection. More specifically, targets containing a coreferential link in the SemEval ABSA 2016 English data set for restaurants were resolved both with AllenNLP's coreference model and manually, and the data sets are publicly available on GitHub[26]. An extensive analysis of implicit targets and the limits of an automatic coreference resolution model are provided, and we show that having sentence-level annotations is beneficial for coreference resolution tasks. Additionally, we confirm that language models improve the results for ABSA tasks.

The results for opinion target extraction indicate that coreference resolution is, in fact, helpful for extracting opinion targets (See Table 9). We observed that most of the best results were obtained with the manually resolved coreference data and that the automatically resolved coreference data worsened the system when trained with RoBERTa. We obtained an F1 score of 0.8168 for simple sequence labelling, which is an improvement of approximately 3 points in the F1 score, and 0.6697 for opinion target extraction with categories, which had no big difference to the original data. Thus, we can conclude that coreference resolution helps to extract opinion targets, but extracting them with their category is a more complex task in which coreference helps less. We report a heightened number of partially correct labels (Figure 21).

For multi-class classification, coreference resolution seems to be more pertinent when there are not several targets in the same sentence. The best results were obtained with automatic and manually resolved coreference, giving us the F1 score of 0.6927, only slightly better than with the original data. For multi-label classification, there seemed to be little improvement, although this method gave better results in general for all the data sets, up to an F1-score of 0.8246 with manually resolved coreference data. In conclusion, we discovered that multi-label classification obtains superior results to multi-class classification, with or without corefrence resolution. Additionally, it seems that coreference resolution is only useful when it was manually resolved. This is due to the fact that since the training data for coreference (OnToNotes) is differs from reviews, our coreference resolution suffers from bad out-of-domain performance. For both types of classification models, the resolution of the ambiguous pronoun *it* seemed to improve the systems.

There are some limits regarding automatic coreference resolution. It was observed that deictic references are especially difficult for coreference, and that locative deixis is relevant for extracting opinions in the restaurant domain. Often customers refer to the restaurant as *here* or *there*, but these were not handled by the coreference resolution model. Also, we could gain more knowledge if other than nominal targets were considered. In the original SemEval 2016 ABSA data set, all targets are nouns, but many opinions are revealed through a verb and an emotional expression, such as *eat well* or *have fun*. Some entities can in fact be events in coference resolution tasks, as explained by Sukthanker et al. (2020).

Generally, opinions that are modelled as entity (noun) and attribute (adjective) are

[26] `https://github.com/rosamariaryh/absa-coref`

easily processed. These follow the annotation model used in the SemEval 2016 ABSA task ENTITY#ATTRIBUTE, where ENTITY is the target and ATTRIBUTE is a polarised expression revealing the aspect. For example, *The food was exceptional.* obtains the aspect category FOOD#QUALITY from the words FOOD#EXCEPTIONAL. However, when the attribute is a verb or adverb, this is not as easily recognised and often NULL target is assigned. For example, *The appetizers we ordered were served quickly* is a simple sentence but, as it does not follow the annotation rule ENTITY#ATTRIBUTE, it is assigned the value NULL. However, if the annotation rule was reconsidered, this could be modelled as SERVE#QUICKLY.

If we train the model with more explicit targets but these do not offer variability, there may not be great improvements in the performance. For example, if we have food or restaurant or staff tagged many times already and we increase the number of these tags, then we might achieve overtagging. We observed in the data that common words like *food* or *place* were over represented in the sequence labelling task, creating false positives. However, if we resolve less coreferential links but they are crucial ones, such as proper names or rare dishes, then this may have a positive impact on the model. Thus, the quality of training data is emphasised in this context as well.

8 Future Work

For future work, similar classification and analysis could be carried out for joint-extraction of targets and sentiment expressions. It has been observed that, although the aspect category SERVICE#GENERAL has a great number of coreferential targets among the implicit targets, the attributes related to them are usually events and other abstract language where it is impossible to pinpoint an adjective, adverb, noun phrase or other element that defines the category and/or sentiment of the target. For example, *However, they do take your cellphone numbers so that you can go hang out somewhere else till they call you up on your cellphone.* has a reference to restaurant staff that could not be resolved due to missing explicit annotation in the document.

Prompting as a means of aspect category detection could be interesting, as it allows to train models with less data (see: Li et al. (2021) and Min et al. (2021a)). Prompting consists of using templates with word gaps that relate to the category instead of labelled training data. This has been proved to be an efficient way to train a model, as it imitates the Masked Language Modelling process in BERT's pre-training step. Due to time limits in the framework, this option was not explored in this thesis.

We have only explored the effect of resolving the implicit targets that have coreferential links (approximately 20% of the implicit targets) with explicit references, which means other types of implicit targets have not yet been analysed and made explicit. For example, for ellipsis, the emotion words and semantic role labelling could be analysed in order to infer the target, and verbal targets could be considered in order to cover more implicit targets. Although these are not always clear entities like nouns, they can still reveal important information about the reviews and the target's category and sentiment.

As a closing remark and in relation with the previous paragraph, it may be worth to reflect on the ethics of manipulating people's reviews for text processing. For coreference resolution, it may not be necessary to ask this question, as the referential links are rather objective, but other implicit targets requiring common world knowledge may include more subjective views. In those cases, it is worth asking, up to what extent can we make explicit targets in people's reviews without changing the original message and point of view.

References

Rodrigo Agerri and German Rigau. Language independent sequence labelling for opinion target extraction. *Artificial Intelligence*, 268:85–95, Mar 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.12.002. URL http://dx.doi.org/10.1016/j.artint.2018.12.002.

Jeremy Barnes, Toni Badia, and Patrik Lambert. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1104.

Jeremy Barnes, Laura Ana Maria Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. Semeval-2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Seattle. Association for Computational Linguistics*, 2022.

David M. Blei, A. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

BrightLocal. Local consumer review survey 2022. 2022. URL https://www.brightlocal.com/research/local-consumer-review-survey/.

Orphée De Clercq and Veronique Hoste. It's absolutely divine! can fine-grained sentiment analysis benefit from coreference resolution? In *CRAC*, 2020.

Orphée De Clercq, Véronique Hoste, and Iris Hendrickx. Cross-domain Dutch coreference resolution. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 186–193, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://aclanthology.org/R11-1026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Xiaowen Ding and Bing Liu. Resolving object and attribute coreference in opinion mining. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 268–276, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1031.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118: 272–299, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.10.003. URL https://www.sciencedirect.com/science/article/pii/S0957417418306456.

Oswald Ducrot. Présupposés et sous-entendus. *Langue Francaise*, 4:30–43, 1969.

Aitor García-Pablos, Montse Cuadros, and German Rigau. W2vlda: Almost unsupervised system for aspect based sentiment analysis, 2017. URL https://arxiv.org/abs/1705.07687.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 1997. Association for Computational Linguistics. doi: 10.3115/976909.979640. URL https://aclanthology.org/P97-1023.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 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, July 2017. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *ACL*, 2019.

Iris Hendrickx and Veronique Hoste. Coreference resolution on blogs and commented news. In Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications*, pages 43–53, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04975-0.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland, September–October 2019. Linköping University Electronic Press. URL https://aclanthology.org/W19-6120.

Véronique Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, 01 2005.

Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. Multi-label few-shot learning for aspect category detection, 2021.

Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, page 755–760. AAAI Press, 2004a. ISBN 0262511835.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004b.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

6280–6285, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1654. URL https://aclanthology.org/D19-1654.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. Bert for coreference resolution: Baselines and analysis. In *EMNLP*, 2019.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

Lauri Karttunen. Discourse referents. In J. D. McCawley, editor, *Syntax and Semantics Vol. 7*, pages 363–386. Academic Press, 1976.

Lauri Karttunen. Presupposition and linguistic context. *Theoretical Linguistics*, 1:181–194, 01 1974. doi: 10.1515/thli.1974.1.1-3.181.

Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-0301.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4):885–916, 12 2013. ISSN 0891-2017. doi: 10.1162/COLI_a_00152. URL https://doi.org/10.1162/COLI_a_00152.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://aclanthology.org/D17-1018.

Kenton Lee, Luheng He, and L. Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL-HLT*, 2018.

Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, and Zhi Yu. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis, 2021.

Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1310. URL https://aclanthology.org/D17-1310.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. Aspect term extraction with history attention and selective transformation, 2018.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. A unified model for opinion target extraction and target sentiment prediction, 2019a.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting bert for end-to-end aspect-based sentiment analysis, 2019b.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. Solving aspect category sentiment analysis as a text generation task, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

Q. Lu, Z. Zhu, and G et al. Zhang. Aspect-gated graph convolutional networks for aspect-based sentiment analysis. In *Appl Intell 51, 4408–4419*, 2021. URL https://doi.org/10.1007/s10489-020-02095-3.

Deon Mai and Wei Emma Zhang. Aspect extraction using coreference resolution and unsupervised filtering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 124–129, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-srw.18.

Merriam-Webster. *Cataphora.* Merriam-Webster. URL https://www.merriam-webster.com/dictionary/cataphora.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey, 2021a. URL https://arxiv.org/abs/2111.01243.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey, 2021b.

Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 339–348, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-1036.

Nicola A. Nicolov, Franco Salvetti, and Steliana Ivanova. Sentiment analysis : Does coreference matter ? 2008.

David E. Over and Paul Grice. Studies in the way of words. *The Philosophical Quarterly*, 40:393, 1989.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135, jan 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL https://doi.org/10.1561/1500000011.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

PewResearch. Most americans rely on their own research to make big decisions, and that often means online searches. 2020. URL https://www.pewresearch.org/fact-tank/2020/03/05/most-americans-rely-on-their-own-research-to-make-big-decisions-and-that-often-mea

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *COLING 2014*, 2014.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495, 2015.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2016 task 5: Aspect based sentiment analysis. In **SEMEVAL*, 2016.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108: 42–49, 2016. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2016.06.009. URL https://www.sciencedirect.com/science/article/pii/S0950705116301721. New Avenues in Knowledge Bases for Natural Language Processing.

Shiva Ramezani, Razieh Rahimi, and James Allan. Aspect category detection in product reviews using contextual representation. 2020.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference, 2021.

Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, Xian wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. Hidden sentiment association in chinese web opinion mining. pages 959–968, 01 2008. doi: 10.1145/1367497.1367627.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.01.010. URL https://www.sciencedirect.com/science/article/pii/S1566253519303677.

Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. DCU: Aspect-based polarity classification for SemEval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 223–229, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2036. URL https://aclanthology.org/S14-2036.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, 2017.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1058. URL https://aclanthology.org/D16-1058.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034721. URL https://aclanthology.org/P99-1032.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.622. URL https://aclanthology.org/2020.acl-main.622.

Hang Yan, Junqi Dai, Tuo ji, Xipeng Qiu, and Zheng Zhang. A unified generative framework for aspect-based sentiment analysis, 2021.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2017. URL `https://aclanthology.org/P18-2017`.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. Useradapter: Few-shot user learning in sentiment analysis. In *FINDINGS*, 2021.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. Movie review mining and summarization. In *CIKM '06*, 2006.