It's absolutely divine! Can fine-grained sentiment analysis benefit from coreference resolution?

Orphée De Clercq
LT3, Ghent University
Groot-Brittannielaan 45
Ghent, Belgium
orphee.declercq@ugent.be

Véronique Hoste
LT3, Ghent University
Groot-Brittannielaan 45
Ghent, Belgium
veronique.hoste@ugent.be

Abstract

While it has been claimed that anaphora or coreference resolution plays an important role in opinion mining, it is not clear to what extent coreference resolution actually boosts performance, if at all. In this paper, we investigate the potential added value of coreference resolution for the aspect-based sentiment analysis of restaurant reviews in two languages, English and Dutch. We focus on the task of aspect category classification and investigate whether including coreference information prior to classification to resolve implicit aspect mentions is beneficial. Because coreference resolution is not a solved task in NLP, we rely on both automatically-derived and gold-standard coreference relations, allowing us to investigate the true upper bound. By training a classifier on a combination of lexical and semantic features, we show that resolving the coreferential relations prior to classification is beneficial in a joint optimization setup. However, this is only the case when relying on gold-standard relations and the result is more outspoken for English than for Dutch. When validating the optimal models, however, we found that only the Dutch pipeline is able to achieve a satisfying performance on a held-out test set and does so regardless of whether coreference information was included.

1 Introduction

In the last two decades, the field of sentiment analysis (SA) has yielded a lot of attention in both academia and commerce (see Liu (2015), Mohammad (2016) or Zhang et al. (2018) for overviews). The attention in SA research has shifted from the coarse-grained detection of the polarity of a given piece of text to the more fine-grained detection of not only polarity, but also the target of the expressed sentiment, as exemplified by the SemEval shared tasks on aspect-based sentiment analysis (Pontiki et al., (2014; 2015; 2016)). In reviews, many references to different aspects of a given product, experience, etc. are made and in a large number of cases, these references are even implicit. Regarding these implicit references, there are two options: either the referent is truly implicit meaning that the aspect can only be inferred from the implied meaning of the sentence, or the referent is an anaphor referring back to an antecedent that was or was not previously mentioned in the review.

While it has been claimed that anaphora or coreference resolution plays an important role in opinion mining to resolve the relationship between the mentioned entities in a given text and across texts (Liu, 2012), it is not clear to what extent coreference resolution actually boosts SA performance, if at all. In this paper, we investigate the potential added value of coreference resolution in the aspect-based SA of restaurant reviews for two languages: English and Dutch. By also manually annotating coreferential links in our data, we measure the incidence of referential links in our review corpus and investigate the upper bound of coreference resolution on SA performance. We reveal that although a certain number of coreferential instances are available in both languages, this does not alter the performance. On the contrary, when relying on automatic coreference resolution systems in both languages, we find that this hampers overall performance.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 11–21,
Barcelona, Spain (online), December 12, 2020.

The remainder of this paper is organised as follows. In the next section, we discuss related work. In Section 3, we explain the datasets that were used for this research and how these have been annotated for both aspect-based sentiment analysis and coreference resolution. Next, in Section 4, we discuss the supervised machine learning classifier that was built and focus on how adding coreference resolution to the pipeline could alter accuracy. We present the results in Section 5, after which we conclude our work in Section 6 and offer prospects for future research.

2 Related work

The large volume of existing work on sentiment analysis from its early days until now can roughly be divided into lexicon-based and machine learning approaches. Lexicon-based methods determine the semantic orientation of a text based on scanning the words occurring in that text while relying on lexicons. Until recently, machine learning approaches were feature-based and applied supervised machine learning algorithms such as Support Vector Machines. With the advent of deep learning end-to-end approaches have also proven to perform well (Zhang et al., 2018).

Both types of approaches have been applied at various levels of a text: the document (Pang et al., 2002), paragraph (O'Hare et al., 2009), sentence (Li et al., 2010), phrase (Wilson et al., 2009) and word (Hatzivassiloglou and McKeown, 1997) level. For each of these levels, coarse-grained as well as fine-grained sentiment analysis can be performed. The latter means that the focus is not only on determining the polarity of a given utterance, but also on the identification of, for example, the source and target of the expressed sentiment (Kim and Hovy, 2006).

In the last decade, a substantial amount of research has been dedicated to target detection for aspect-based sentiment analysis (Pontiki et al., 2014). This task focuses on the detection of all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer. Following the SemEval task description, aspect-based sentiment analysis can be decomposed into three subtasks: aspect term extraction, aspect term aggregation or classification and aspect term polarity estimation. The focus of the research presented here is on the second one.

The idea is to predict several predefined and domain-specific categories, i.e. a multiclass classification task. The two systems achieving the best results on this individual subtask in SemEval 2015 Task 12 both used classification to this purpose, respectively individual binary classifiers trained on each possible category which are afterwards entered in a sigmoidal feedforward network (Toh and Su, 2015) and a single Maximum Entropy classifier (Saias, 2015). When it comes to feature engineering, especially lexical features in the form of bag-of-words such as word unigrams and bigrams (Toh and Su, 2015) or word and lemma unigrams (Saias, 2015) and lexical-semantic features in the form of clusters learned from a large corpus of reference data (Saias, 2015) were used.

Since then, these benchmark SemEval datasets have been used many times to train and test neural models yielding state-of-the-art results on both this second subtask and end-to-end aspect-based sentiment analysis (Do et al., 2019) However, many methodologies start from he assumption that the target of a given polarity is explicitly lexicalized, which is certainly not always the case as people often use shorter or alternative linguistic structures, such as anaphors to refer to previously mentioned elements.

Many survey studies on sentiment analysis have claimed that the recognition of coreference is crucial for successful (aspect-based) sentiment analysis (Liu, 2012; Feldman, 2013). Stoyanov and Cardie (2006) were the first to use coreference resolution features to determine which mentions of opinion holders refer to the same entity. Early research in incorporating basic coreference resolution in sentiment analysis was conducted by Nicolov et al. (2008), who investigated how to perform sentiment analysis on parts of the document around topic terms. They demonstrated that using a proximity-based sentiment analysis algorithm can be improved by about 10%, depending on the topic, when using coreference to augment the focus area of the algorithm. The work by Kessler and Nicolov (2009), though its main focus is on finding which sentiment expressions are semantically related, provided some valuable insights in the necessity of coreference as they found that 14% of the targets expressions that had been manually labeled in their corpus were expressed in the form of pronouns. Ding and Liu (2010) introduced the problem of entity and aspect coreference resolution and aimed to determine which mentions of

entities and/or aspects a certain pronoun refers to, taking a supervised machine learning approach. Their system learns a function to predict whether a pair of nouns is coreferent, building coreference chains based on feature vectors that model a variety of contextual information about the nouns. They also added two opinion-related features, which implies that they used sentiment analysis for the purpose of better coreference resolution rather than the other way around. A similar coreference resolution methodology was used by Zhao et al.(2015) to link target aspects to target objects. However, to our knowledge not much qualitative research has been performed investigating to what extent the availability of coreference information can actually help aspect-based sentiment analysis. This is the exact aim of this paper, we zoom in on the task of aspect category classification and investigate whether including coreference information prior to classification is useful.

3 Datasets and annotation

For our experiments, we rely on datasets comprising restaurant reviews in two languages, namely English and Dutch. Both datasets were released in the framework of SemEval: the English data (350 reviews) was released for the 2015 competition (Pontiki et al., 2015) and the Dutch data (400 reviews) for a rerun of this competition in 2016 (Pontiki et al., 2016).

3.1 ABSA annotation

All English and Dutch restaurant reviews were annotated following the SemEval ABSA guidelines[1]. Every review was split into sentences and a sentence was only annotated with aspect terms and categories when a polarity was expressed in the sentence. In total, 1,702 English (85%) and 1,767 Dutch (76%) sentences were found to be opinionated and further annotated with targets, aspect categories (i.e. Ambience, Drinks, Food, Location, Restaurant and Service) having different attributes (i.e. General, Prices, Quality, Style & Options and Miscellaneous) and polarity. Important for the research presented here is to note that a distinction was made between explicit and implicit targets.

Whenever there was an explicit target, the span of the terms evoking that target was included in the annotation; implicit targets were added as 'NULL' targets. As a consequence, pronouns are not annotated as separate targets, even if they refer to an explicit target. Instead, those pronouns, together with other aspects that are referred to implicitly, are added as 'NULL' targets, which are then further annotated with aspect categories and polarities. In Table 1, we give an overview of how many different aspect categories are available in both datasets, together with the number of implicit targets. We observe that 623 out of the 2,499 annotated aspect categories for English (24.9%) and 773 out of the 2,445 for Dutch (31.6%) are implicit or 'NULL' targets.

Main	**Attribute**	**Total**		**Implicit**	
		EN	DU	EN	DU
Ambience	General	260	240	28	56
Drinks	Prices	20	23	0	3
	Style & Options	32	38	0	4
	Quality	46	68	3	3
Food	General	1	15	0	4
	Prices	85	54	18	19
	Style & Options	133	209	19	27
	Quality	852	675	86	123
Location	General	28	34	6	7
Restaurant	General	416	437	233	296
	Prices	83	43	57	33
	Miscellaneous	100	26	51	12
Service	General	443	583	122	186
Total		2499	2445	623	773

Table 1: Total number of annotated aspect categories and implicit targets

[1] http://alt.qcri.org/semeval2016/task5/data/uploads/absa2016_annotationguidelines.pdf

3.2 Coreference annotation

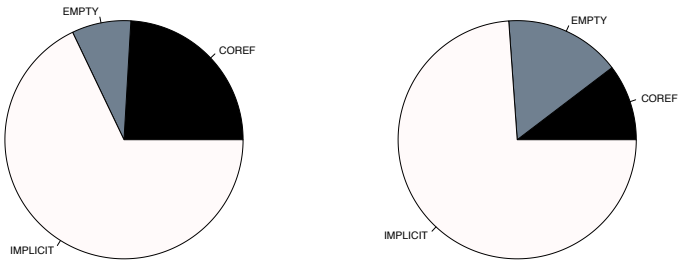
We manually annotated each implicit or 'NULL' target by indicating whether this implicit target was clearly referential (i.e. an anaphor), whether the antecedent was also mentioned in the review or whether none of both applied. We only indicated coreferential relations constituting an identity (and thus not a part-of, etc.) relation between the anaphor and the antecedent. The following three examples exemplify this annotation procedure.

1. *I tend to judge a sushi restaurant by [its sea urchin]. [It] melted in my mouth and was perfect.*
   It = anaphor
   its sea urchin = antecedent

2. *This place is incredibly tiny. [They] refuse to seat parties of 3 or more on weekends.*
   They = anaphor
   antecedent not mentioned; 'staff' is implied

3. *Can't wait wait for my next visit.*
   No anaphor, no antecedent

The pie charts below illustrate the subdivision of these additional annotations in our datasets: for each 'NULL' target we indicated whether the implicit target was referential, and if so, whether the antecedent was mentioned (COREF) or not (EMPTY). If there was no referential relation, we labelled it as IMPLICIT.



English Dutch

Figure 1: Pie charts visualizing the implicit target distribution in our datasets.

In both languages the vast majority is labelled as implicit. Regarding the usage of referential pronouns we observe a different tendency: in English, an anaphor is more frequently preceded by an antecedent within one review, whereas for Dutch more empty anaphors are included. Percentagewise, we see that in our English dataset 20.7% of the implicit targets are referential pronouns for which the antecedent can be discovered within the same review, whereas for Dutch this is only true for 10.34% of the implicit reviews. When performing coreference resolution prior to classification into aspect categories, we can therefore assume that this technique will be more successful for English.

4 Experimental setup

The focus of our experiments is on the task of aspect category classification. This is a fine-grained classification task requiring a system to grasp subtle differences between various main–attribute categories (e.g. *Food–General* versus *Food–Prices* versus *Food–Quality* versus *Food–Style&Options*). Moreover, as previously explained, reviewers refer to the various aspects of a restaurant in both an explicit and implicit manner. Especially those implicit targets are challenging. This is why we will investigate whether including coreference information prior to classification is useful. We envisage two experimental settings: a setting where coreferential anaphor–antecedent pairs were not derived beforehand and one where they were. In the latter setting, both gold-standard and automatically-derived coreference relations were used in order to investigate the true upper bound of incorporating this type of information.

We relied on gold-standard explicit and implicit targets for all experiments. As experimental data we employed the same train and test splits of the SemEval shared tasks on ABSA (Pontiki et al., 2016), see Table 2.

	ENGLISH		DUTCH	
	train	**test**	**train**	**test**
# targets	1654	845	1843	602
# implicits	375	248	563	210
# explicits	1279	597	1280	392

Table 2: the overall number of targets and the number of implicit and explicit targets in the datasets.

4.1 Information sources

As a baseline, we derived bag-of-words token unigram features of the sentence in which a target occurs in order to represent some of the lexical information present in each of the categories. In bag-of-words representations, each feature corresponds to a single word found in the training corpus. Besides these lexical features, features in the form of clusters derived from a large domain-specific reference corpus have also proven useful (Toh and Su, 2015; Toh and Su, 2016). Given the lack of such reference corpora for Dutch, we decided to link mentions of concepts and instances to either semantic lexicons like WordNet (Fellbaum, 1998)(English) or Cornetto (Vossen et al., 2013) (Dutch) , and to a Wikipedia-based knowledge base (Hovy et al., 2013) such as DBpedia (Lehmann et al., 2013).

This led to the creation of a set of lexico-semantic features. Six WordNet features were derived, each representing a value indicating the number of (unique) terms annotated as aspect terms from that category that (1) co-occur in the synset of the candidate term or (2) which are a hyponym/hypernymof a candidate term in the synset. Furthermore, we identified concepts in DBpedia by processing each target with DBpediaSpotlight (Mendes et al., 2011). Next, categories for each concept were created, corresponding to the categories in Wikipedia. To that end, we extracted all direct categories for each concept (`dcterms:subject`), and added the more general categories with a maximum of two levels up in the hierarchy (`skos:broader`). This process is illustrated in Figure 2. The whole process, comprising the annotation with DBpedia Spotlight and the extraction of categories, was performed using the RapidMiner LOD Extension (Paulheim and Fürnkranz, 2012).

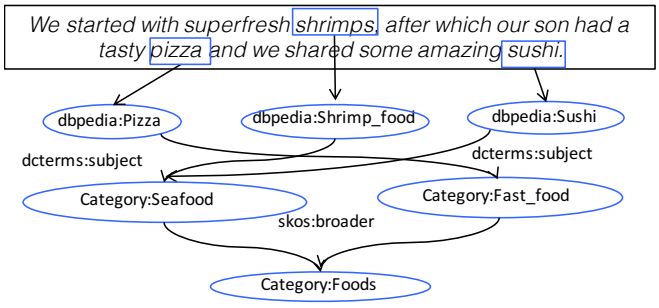


Figure 2: Example sentence in which targets are semantically enriched using DBpedia.

4.2 Coreference resolution

As all implicit aspect mentions and pronouns referring to aspects had been annotated as 'NULL' targets it was impossible to derive lexico-semantic features for these instances. However, because coreference information was added to these aspects, we hypothesized that for certain 'NULL' targets these features can actually be derived. In other words, a coreferential relation between an anaphor – pronoun – and an antecedent constituting an aspect term in itself should enable us to derive additional semantic information.

For the research presented here, we explored the added value of incorporating coreference information by including it as a separate processing step before the feature extraction. Crucial for this step is that the coreference resolution is highly accurate, since an anaphor–antecedent mismatch can also lead to a semantic information mismatch. To this purpose, we relied on existing systems in both languages: the deterministic Stanford Coreference Resolver (Lee et al., 2013) for English and the COREA system for Dutch (De Clercq et al., 2011).

The Stanford system is a rule-based system that includes a total of ten rules (or "sieves") for entity coreference, such as exact string match and pronominal resolution. The sieves are applied from highest to lowest precision, each rule adding coreference links. The COREA system is a mention-pair system (Hoste, 2016) that recasts the coreference resolution problem as a classification task: a classifier is trained to decide whether a pair of noun phrases or mentions is coreferential of not. In other words, resolving anaphor m_j can be viewed as the task of finding the mention m_i that maximizes the probability of the random variable L:

$$argmax_{m_i} P(L|m_j, m_i)$$

In the mention-pair model, each pair of NPs is represented by a feature vector containing distance, morphological, lexical, syntactic and semantic information on both NPs and the relation between them. The goal of the feature information is to enable the machine learner to distinguish between coreferential and non-coreferential relations, and for example to resolve that *it* in example 2 does not refer to *a sushi restaurant*, nor to *sushi rose*, but to *its sea urchin*. After this binary classification, a second step, a separate clustering mechanism is used to coordinate the pairwise classification decisions and to build so-called 'coreference chains'.

As we also manually annotated each 'NULL' aspect term constituting an anaphor–antecedent relation, we were able to asses the upper bound of incorporating coreference information for this task.

4.3 Optimization

Our main interest is to explore whether, and if so, how the subtask of aspect category classification, which typically relies on shallow lexical characteristics and some incorporation of semantic information, can benefit from incorporating coreference information. This is done by including coreference resolution as a preprocessing step prior to classification. To this purpose, the experiments on the training data were split in a setting where coreference relations are not derived beforehand (Setting A) and one where they are (Setting B). In the latter setting, a comparison is also made between automatically-derived and gold standard coreference information in order to assess the true upper bound.

Ten-fold cross validation experiments are conducted on the training set using LibSVM[2], version 3.17 (Chang and Lin, 2011) and we evaluate the results using accuracy as performance metric.

In both settings, we used genetic algorithms to derive the optimal feature combinations. Since each machine learning algorithm's performance is inherently dependent of the different parameters that are used, we performed a joint optimization in two different scenarios. We allow 100 generations and set the stopping criterion to a best fitness score (accuracy) that remained the same during the last five generations. Our search starts from a population of 100 individuals and all optimization experiments are performed using the Gallop toolbox (Desmet and Hoste, 2013).[3] In the first scenario (featgroups), we perform hyperparameter and feature group selection using the three feature groups we have available (i.e. bag-of-words, WordNet and DBpedia) and allow variation in LibSVM's hyperparameters. In the second scenario (indfeats), we perform hyperparameter selection and allow individual feature selection among the lexical-semantic (WordNet and DBpedia) features. The bag-of-words features are kept together as a group.

In a final experiment, the optimal settings emerging from the experiments on the training data in Setting A and B are tested on the held-out test set.

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[3] For more information we refer to (2016) where similar experiments were performed for the task of readability prediction.

5 Results

In Setting A, coreference resolution is not performed prior to classificiation, so only the explicit aspect terms are targeted. In setting B, coreference resolution is included as an additional processing step prior to classification. Having coreference information available should allow us to derive additional semantic information for those 'NULL' targets constituting an anaphor–antecedent pair. We differentiate between a setup where we incorporate this information assuming we have a perfect coreference resolution system (GOLD), i.e. using gold-standard coreferential links, and a setup where coreference relations have been resolved automatically (AUTO). Coreference resolution as an additional processing step prior to classification. The results, expressed in accuracy, are presented in Table 3.

	ENGLISH		DUTCH	
	Joint optimization		Joint optimization	
	featgroups	indfeats	featgroups	indfeats
SETTING A	67.17	67.23	62.94	63.16
SETTING B (GOLD)	67.96	**68.20**	62.78	**63.59**
SETTING B (AUTO)	67.07	67.23	60.77	60.88

Table 3: Results of cross-validation experiments on the training data.

Overall, we observe that, when using gold information, the results increase in both languages, an accuracy of 68.20 for English and one of 63.59 for Dutch. This indicates that including coreferential links between anaphor–antecedent pairs is beneficial. If we resolve coreference automatically, however, we see that our results decrease or remain on par with the results without coreference.

From the above-mentioned results, it can also be concluded that the added value of including coreference information is not outspoken. When relying on coreference resolution systems, the performance mostly deteriorated mainly because wrong antecedents have been linked to anaphors, causing erroneous lexical-semantic features. However, our results also revealed that incorporating gold-standard anaphor-antecedent relations leads to the best overall scores in both languages after jointly optimizing LibSVM's hyperparameters and performing individual feature selection. If we compare these scores to the best individual scores achieved in setting A, we observe that the difference is more outspoken for English, which confirms our hypothesis. In the next section we will analyse whether incorporating coreference information also meant that the lexical-semantic features were considered more important.

5.1 Feature informativeness in both settings

In order to discover the added value of the lexical semantic features, we compared both optimal settings and will discuss which hyperparameters, and especially which lexical-semantic features were selected in both languages. Because, at the end of a GA optimization run, the highest fitness score may be shared by multiple individuals having different optimal feature combinations or parameter settings, we also considered runner-up individuals to that elite as valuable solutions to the search problem. This is why the features are visualized using a color range: The closer to blue, the more this feature group was turned on and the closer to red, the less important the feature group was for reaching the optimal solution. The numbers within the cells represent the same information but percentagewise.

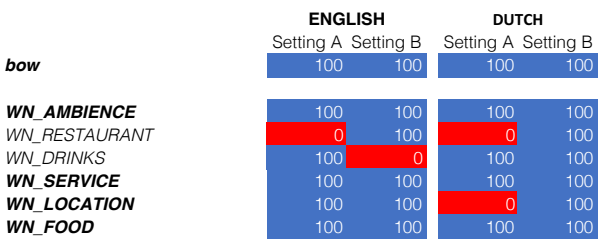


Figure 3: Where the bag-of-words features (bow) selected and which WordNet features (WN) were selected in the optimal setting.

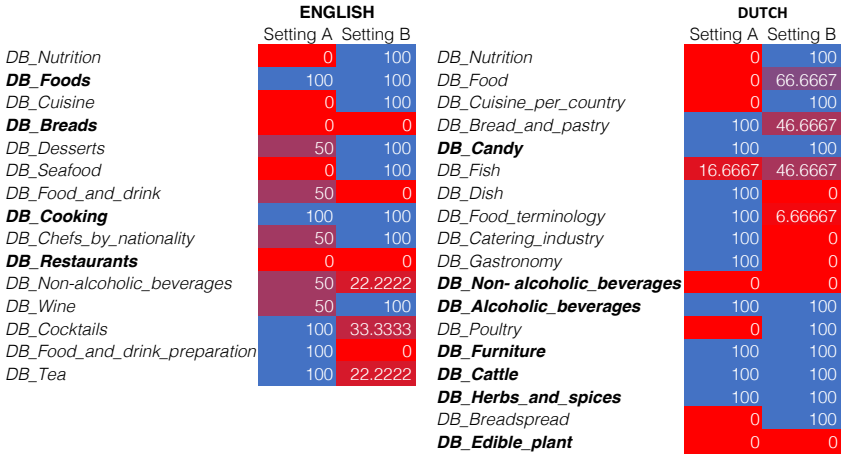


Figure 4: Which DBpedia features (DB) were selected in the optimal settings.

As can be derived from Figure 3, we observe that for both languages the bag-of-words features are crucial and always selected. Regarding the WordNet features, for English in both settings five features are selected, though not the same five. In the setting without coreference information, the feature related to the main aspect category restaurant is not selected. Whereas, in the other setting the same goes for the feature related to the aspect category drinks. For Dutch, we observe that all WordNet features are turned on when (gold-standard) coreference information has been included prior to classification.

For the DBpedia features, listed in Figure 4, there are differences between both languages. For English, we notice that only four out of the fifteen features remain unchanged in both settings, these are indicated in bold. Overall, we observe that more DBpedia features are turned on in the setting with coreference information, i.e., eight versus five features. For Dutch, seven out of the eighteen features remain unchanged and though there is a shift as to which features are selected in the optimal setting with coreference information, we see that only nine feature groups are turned on, compared to ten that were turned on in the optimal setting without coreference information.

For both languages we can conclude that including semantic information in the form of lexical-semantic features is important as a large number of these features are being selected in the optimal settings. When we look at the optimal setting with coreference information, we observe that especially for English more DBpedia features are being turned on.

5.2 Testing optimal models on held-out test sets

In a final round of experiments, the two optimal models were tested on the held-out test sets. The results are presented in Table 4.

	Train	Held-out test
Optimal EN model Setting A	67.23	57.75
Optimal EN model Setting B	**68.20**	56.92
Optimal DU model Setting A	63.16	**66.42**
Optimal DU model Setting B	63.59	**66.42**

Table 4: Comparison of the optimal results on the training data and of the held-out experiments

Though the distribution between the explicit and implicit targets does not differ between the train and test sets in both languages (Table 2), we do observe different results. For English, there is a dramatic drop in performance on the held-out test set, for setting A we achieve an accuracy of 57.75% and for setting B an even lower one of 56.92%. With these results, we are far below the best performing system at the SemEval 2016 task (Pontiki et al., 2016), but, as stated previously, we only relied on a limited amount of information sources because of comparison purposes with Dutch. Contrary to our expectations, coref-

erence information, even when added as gold standard anaphor-antecedent pairs, does not help to reach a better performance. For Dutch, on the other hand, we achieve an accuracy of 66.42 in both settings, which is three points higher than the best accuracy scores on our training set. This result is also almost ten points higher than the best result achieved on this dataset at the SemEval 2016 task (Pontiki et al., 2016). However, these results also indicate that on our held-out test set there is no difference between the accuracy obtained with or without adding gold-standard coreference relations prior to classification.

Surprised with these outcomes regarding the added value of coreference information, especially for English where the results even deteriorated, we inspected the subdivision of the implicit aspects in both held-out test sets. We found that in the English set 240 out of the 248 implicit targets were truly implicit and that out of the eight referential anaphors, only four referred back to an antecedent within the same review. In the Dutch test set, 154 out of the 210 were truly implicit and out of the 56 referential anaphors, thirty instances constituted an anaphor-antecedent pair within the same review.

6 Conclusion

The objective of this research was to investigate to what extent coreference resolution can boost sentiment analysis performance. Our focus was on aspect-based sentiment analysis of English and Dutch restaurant reviews and more specifically the task of classifying aspect terms into predefined aspect categories. We worked with two datasets that were released and annotated in the framework of SemEval. Working with these datasets, we found that people often refer to aspect terms implicitly in both languages (24.9% in English versus 31.6% in Dutch).

This is why we investigated whether including coreference information prior to classification would be useful for pinpointing those implicit aspect terms constituting a referential relation with an antecedent. We manually annotated coreferential relations in both datasets and observed a different tendency in both languages. In English, an anaphor is frequently preceded by an antecedent within one review, whereas for Dutch the anaphors more frequently refer to extra-linguistic entities which are not explicitly mentioned in the review. When exploiting coreferential information in an aspect-based sentiment analysis pipeline, we therefore hypothesized that this would be more successful for English than for Dutch.

To investigate this, experiments were conducted in two different settings: a first setting where coreferential anaphor-antecedent pairs were not derived beforehand and a second setting where they were. In the latter setting, both gold-standard and automatically-derived coreference relations were used in order to investigate the true upper bound of incorporating this type of information. Our classifier relied on a combination of lexical (bag-of-words) and lexical-semantic information in the form of WordNet (Fellbaum, 1998) and DBpedia (Lehmann et al., 2013) features. Besides exploring the added value of coreference information, we also used a wrapper-based genetic algorithm optimization approach to optimize our classifiers and get more insights into which features are most important.

The results reveal that resolving coreferential relations prior to classification is beneficial in both settings in a setup where both the hyperparameters and individual features are jointly optimized. However, this is only the case when relying on gold-standard coreferential information and the result is more outspoken for English (from 67.23% to 68.20%) than for Dutch (from 63.16% to 63.59%). Regarding the selected features in the optimal models we could conclude that lexical bag-of-words are necessary to include and that including semantic information in the form of lexical-semantic features is also important. Comparing the optimal setting with and without performing coreference resolution prior to classification, we observe that especially for English more DBpedia features are being turned on.

In a final set of experiments we envisaged to validate these findings by testing the optimal models on a held-out test set in both languages. For English this led to poor results whereas for Dutch we were able to achieve a satisfying performance of 66.42%. In both languages, however, it was no added value to have gold-standard coreference information available before classification.

Though the results now seem to indicate that coreference information is not necessary to include in a fine-grained sentiment analysis pipeline, it will be interesting to corroborate these findings on larger datasets and on data coming from different domains. Now the focus was on resolving anaphor-antecedent pairs within one review, but in reality coreference also appears across texts which offers other interesting

prospects for future research.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27.

Orphée De Clercq and Veronique Hoste. 2016. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *COMPUTATIONAL LINGUISTICS*, 42(3):457–490.

O. De Clercq, I. Hendrickx, and V. Hoste. 2011. Cross-domain Dutch coreference resolution. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP-2011)*, pages 186–193.

B. Desmet and V. Hoste. 2013. Fine-grained Dutch named entity recognition. *Language Resources and Evaluation*, pages 307–343.

X. Ding and B. Liu. 2010. Resolving Object and Attribute Coreference in Opinion Mining. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010)*, pages 268–276.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272 – 299.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

C. Fellbaum. 1998. *WordNet: an Electronic Lexical Database*. MIT Press.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL-1997)*, pages 174–181.

Veronique Hoste. 2016. The mention-pair model. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora resolution : algorithms, resources and applications*, pages 281–295. Springer-Verlag.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Jason S. Kessler and Nicolas Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *The 3rd Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (ICWSM-2009)*, pages 90–97.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST-2006)*, pages 1–8.

H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sôren Auer, and Christian Bizer. 2013. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6:167–195.

Binyang Li, Lanjun Zhou, Shi Feng, and Kam-Fai Wong. 2010. A unified graph model for sentence-based opinion retrieval. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 1367–1375.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics-2011)*, pages 1–8.

Saif M. Mohammad. 2016. Challenges in sentiment analysis. In D. Das, E. Cambria, and S. Bandyopadhyay, editors, *A Practical Guide to Sentiment Analysis*. Springer.

N. Nicolov, F. Salvetti, and S. Ivanova. 2008. Sentiment analysis: Does coreference matter? In *Proceedings of the Symposium on Affective Language in Human and Machine*, pages 37–40.

Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. 2009. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st International Conference on Information and Knowledge Management Workshop on Topic-sentiment Analysis for Mass Opinion (TSA-2009)*, pages 9–16.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86.

Heiko Paulheim and Johannes Fürnkranz. 2012. Unsupervised Generation of Data Mining Features from Linked Open Data. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS-2012)*, page 31.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

José Saias. 2015. Sentiue: Target and aspect based sentiment analysis in SemEval-2015 Task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 767–771, June.

Veselin Stoyanov and Claire Cardie. 2006. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 336–344.

Zhiqiang Toh and Jian Su. 2015. NLANGP: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 496–501, June.

Zhiqiang Toh and Jian Su. 2016. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 282–288.

P. Vossen, I. Maks, R. Segers, H. van der Vliet, M.F. Moens, K. Hofmann, E. Tjong Kim Sang, and M. de Rijke. 2013. Cornetto: a lexical semantic database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, pages 165–184. Springer.

T. Wilson, J. Wiebe, and P. Hoffman. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis : A survey.

Y. Zhao, B. Qin, T. Liu, and W. Yang. 2015. Aspect-Object Alignment with Integer Linear Programming in Opinion Mining. *PLOS One*, 10(5).