Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2816–2825
Marseille, 20-25 June 2022

Building Sentiment Lexicons for Mainland Scandinavian Languages Using Machine Translation and Sentence Embeddings

Peng Liu*, **Cristina Marco***,†, **Jon Atle Gulla**
Department of Computer Science, NTNU, Trondheim, Norway
{peng.liu,jon.atle.gulla}@ntnu.no,cristinasmarco@gmail.com

Abstract

This paper presents a simple but effective method to build sentiment lexicons for the three Mainland Scandinavian languages: Danish, Norwegian and Swedish. This method benefits from the English Sentiwordnet and a thesaurus in one of the target languages. Sentiment information from the English resource is mapped to the target languages by using machine translation and similarity measures based on sentence embeddings. A number of experiments with Scandinavian languages are performed in order to determine the best working sentence embedding algorithm for this task. A careful extrinsic evaluation on several datasets yields state-of-the-art results using a simple rule-based sentiment analysis algorithm. The resources are made freely available under an MIT License.

Keywords: Sentiment lexicon, Scandinavian languages, Machine translation, Sentence embedding

1. Introduction

Ubiquitous and cognitive artificial intelligence is becoming a reality that is shaping our world. By interacting through voice with conversational systems, it is now possible to make purchases online or follow a recipe while you are in the kitchen. Besides, knowledge derived from continuous text data sources coming from social media or online news sources plays an increasingly important role in political campaigns, financial analysis, or analysis of medical records. The fine-grained analysis of sentiment and emotion is becoming essential for all these tasks.

Automatically extracting the positive or negative orientation that a passage expresses toward some targets, so-called sentiment analysis or opinion mining, is one of the fundamental tasks of text categorization. As these opinions or sentiments are written in natural languages and considering the lack of training corpora in several languages, one of the most important tools needed to do such analysis is sentiment lexicons, which contain lists of positive, negative and, in some cases, neutral words. Despite the fact that English resources exist to do this task, sentiment lexicons or training corpora in other languages are often not easily available.

To alleviate the problem raised above, we propose a simple but effective method to develop large general sentiment lexicons for the three Mainland Scandinavian languages, Danish, Norwegian and Swedish. Specifically, to create these resources, machine translation and sentence embeddings methods are used to map the sentiment information from an already available English resource to a thesaurus in one of the target languages. Machine translation is used to bridge this lexicon into several languages. In particular, we focus on experimenting with Danish, Swedish and Norwegian. Because there are a handful of evaluation datasets for these three under-resourced languages, we construct four new datasets based on Tripadvisor reviews as a new benchmark to evaluate the generated resources. The development of a pool of basic natural language processing resources for Scandinavian languages is essential in order to build competitive and state-of-the-art artificial intelligence pipelines for these languages. The resources presented in this paper are made freely available under the MIT License[1].

2. State of the Art

A common approach to sentiment analysis is to use supervised learning. Given an input dataset annotated with relevant sentiment information, the goal of the supervised algorithm is to learn how to map from a new observation to the correct sentiment. In many cases, there are insufficient labelled training data to train accurate classifiers. To deal with this, it is possible to use sentiment lexicons containing a list of positive and negative words as features. There are many sentiment lexicons for English, such as the General Inquirer (Stone et al., 1966), Linguistic Inquiry and Count (LIWC) (Pennebaker et al., 2001), the Opinion Lexicon of Hu and Liu (Hu and Liu, 2004), the MPQA Subjectivity Lexicon (Wilson et al., 2005) and SentiWordnet (Baccianella et al., 2010), among others.

Many sentiment lexicons are created manually. The disadvantages of this method are clear: it is expensive and time-consuming to build resources by hand. MPQA and LIWC were built by human annotators (Pennebaker et al., 2001; Wilson et al., 2005). Alternatively, these resources can be learned automatically or semiautomatically. Semi-supervised approaches to

*Equal contribution.
†This work was done prior to the author joining Amazon.

[1] https://drive.google.com/file/d/10zjHxJQt5Ev86N8pOxDvRKr9GiW6LmkM/view?usp=sharing

sentiment lexicon learning often start from a seed of polarity words like *good* or *bad*, and then find ways to enlarge the lexicon by labelling each word based on its similarity to the two sets of seeds, see (Hatzivassiloglou and McKeown, 1997) and, more recently, (Turney and Littman, 2003), (Hamilton et al., 2016) or (An et al., 2018). Finally, there are also some approaches that make use of a thesaurus like WordNet, containing word synonyms and information about the different senses of a word (Kim and Hovy, 2004; Hu and Liu, 2004). For example, SentiWordnet was built by assigning polarity information to each of WordNet senses (Baccianella et al., 2010). In this approach, polarity is assigned to entire synsets (sets of synonym words).

However, it is more challenging for languages other than English to find sentiment lexicons. There are two main strategies that have been used to build sentiment lexicons in these languages. The first strategy leverages machine translation to directly transfer polarity information from English resources to the target language. The second strategy does sentiment analysis directly on the target language by using existing linguistic resources either in English or from the same target language. The first strategy has proven to be inefficient so far. For example, the work presented in (Mihalcea et al., 2007) for Romanian, (Wan, 2008) for Chinese and (Tsakalidis et al., 2018) for Greek has shown that simply translating a subjectivity or polarity lexicon in a target language does not create a high accuracy resource. As will be shown later, the present paper challenges these results, since one of our findings is that translation can work sufficiently well provided good machine translators exist for the given language combinations. Most approaches in the second strategy try to benefit from the polarity information present in the English resources. Others, in contrast, make use of already existing native resources, either dictionaries or corpora. For example, Perez-Rosas et al. (2012) use a manually annotated lexicon from English, the Opinion Finder lexicon, to enforce SentiWordNet (Esuli and Sebastiani, 2006) based constraints. This information is transferred to the Spanish WordNet by benefiting from the mappings present in the multilingual WordNet.

Our approach is similar to the one in (Perez-Rosas et al., 2012) as we benefit from the English SentiwordNet resource and the WordNet in one of the target languages. However, the mapping of sentiment weights from the English to the Scandinavian resource is more challenging in our case as the cross-lingual links between both resources were not available at the time of writing. In turn, this provided us with the opportunity of making our approach more general and generally applicable to other thesaurus-type of resources.

2.1. Sentiment Resources for Scandinavian Languages

To date, the approaches proposed to deal with Danish are based mainly on semi-supervised methods building on the top of engineered features from social media corpora in combination with a corrupt model to avoid overfitting of the machine learning model (Elming et al., 2014). The Sentida lexicon is one of the most substantial contributions to the field of Danish sentiment analysis so far (Lauridsen et al., 2019). This lexicon consists of 5263 words that have been manually rated on a discrete interval scale ranging from -5 (negative) to +5 (positive). Recently, Pedersen et al. (2021) proposed a collection of Danish lexical semantic resources, which comprises a Danish wordnet, the Danish FrameNet Lexicon and the Danish Sentiment Lexicon. The Norwegian and Swedish languages present diverse cases as far as the lack of sentiment resources is concerned. At the time of writing, a newly sentiment lexicon was published for Norwegian (Barnes et al., 2019). This resource contains 14,839 negative words and 6,103 positive words. It was created firstly by automatically translating the English lexicon of (Hu and Liu, 2004), and, secondly, by manually correcting the translated resource in order to improve its quality. Besides, there has recently been released a Norwegian Review corpus that can be used for the purpose of evaluating sentiment resources (Velldal et al., 2018).

In contrast, there are several sentiment resources to deal with the Swedish language. Specifically, there are three available lexicons: (Rosell and Kann, 2010), (Nusko et al., 2016) and (Rouces et al., 2018a). Rosell and Kann (2010) present a Swedish lexicon containing 1,349 words developed by using random walks over a graph of synonyms and a set of seeds of four positive and four negative words. Nusko et al. (2016) propose a tree traversal method on SALDO, starting with six seeds. The resulting sentiment lexicon has 2,133 entries with a precision of 71% computed on the basis of a manual evaluation of 100 words from this lexicon. More recently, Rouces et al. (2018a) report their best results when using word embeddings, but still in the range of 65% for the positive and negative classes.

3. The Resources

Table 1 summarizes the new resources for Danish, Swedish and Norwegian. As can be seen from the table, these resources contain around 33 and 35 thousand synsets and senses, where around 10% are positive and 10% negative, the rest being neutral. These resources contain additional valuable information, such as the gloss, the identity number of the corresponding synset in the English resource, the part-of-speech and the sense number and identity number of DanNet. The method used to build these resources is described in Section 5.

4. Datasets

Two datasets were used in our experiments: SentiWordnet and DanNet. SentiWordNet is a lexical resource in which each WordNet synset is associated with three numerical scores Obj(s), Pos(s) and Neg(s),

SL	Synset	Senses	Positive	Negative
Danish SWN	33,251	35,718	10.5%	10.7%
Swedish SWN	33,221	35,032	10.5%	10.7%
Norwegian SWN	33,224	35,036	10.5%	10.7%

Table 1: Summary of the generated Danish, Swedish and Norwegian sentiwordnets.

	Synsets	Senses	Positive	Negative
SentiWordNet	117,374	206,470	11%	12%
DanNet	65,583	74,718	-	-

Table 2: Summary of SentiWordNet and DanNet lexical resources

describing how objective, positive, and negative the English terms or wordsenses contained in the synset are[2]. WordNet synsets are nouns, verbs, adjectives or adverbs grouped into sets of cognitive synonyms, each expressing a distinct concept. For example, the wordsense *unable#1* in this lexicon has a 0.75 negative score, and the wordsense *able#1* has a 0.125 positive score. DanNet is the so-called WordNet for Danish[3]. Table 2 summarizes the number of synsets and senses in each of these resources. As can be seen from this table, SentiWordnet is considerably larger than DanNet. The former contains almost 120 thousand sets of synonyms, whereas the latter includes 65 thousand. Besides, there are more senses per synset in SentiWordNet than DanNet. The average number of senses per synset is 1.76 for SentiWordnet, whereas only 1.14 for DanNet. The last columns in this table indicate the percentage of mostly positive and negative synsets in SentiWordNet. It is interesting to observe that most synsets in this thesaurus show a neutral orientation, whereas only a quarter suggests a positive or negative orientation. These numbers are similar to the ones observed in Table 1 about our resources.

5. Method

In this paper, we propose a method to automatically create sentiment lexicons for Scandinavian languages by employing a score-based polarity lexicon for English, SentiWordnet (Baccianella et al., 2010) and the Danish version of WordNet, DanNet (Pedersen et al., 2009). Despite the fact that there are WordNets of Norwegian and Swedish[4], they do not have glosses of the mapping words, and we cannot measure the similarities with them.

Figure 1 illustrates this method. Automatic translation is used to bridge both resources, the English Senti-

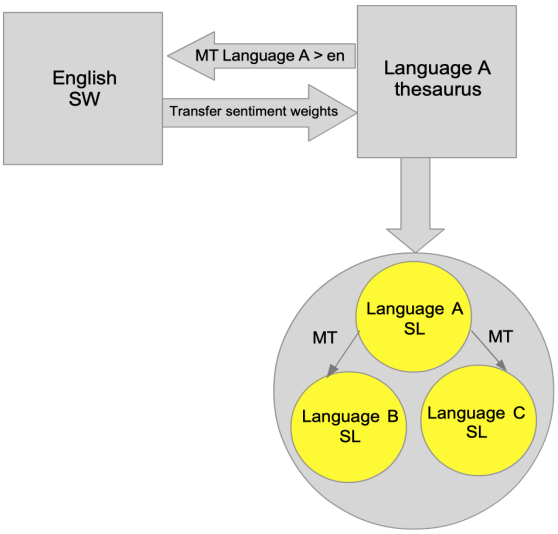

Figure 1: Method to build sentiment lexicons (SL) for under-resourced languages.

Wordnet and the thesaurus in Language A. The mapping between both resources is done by translating all senses and glosses (or definitions) included in Language A into English. Then, a direct mapping can be done in those cases in which there is a direct match between the wordsense from the English resource and the one in Language A. If there are several similar senses in both resources, a similarity measure is computed on the glosses in order to align both resources based on the most similar meaning. Once aligned, the polarity weights present in the English resource are mapped into the resource in Language A. After this new sentiment lexicon for Language A is created, the resource is automatically translated into other languages, say B and C, in order to generate additional sentiment lexicons. Assuming that a higher degree of similarity exists between these languages and accurate translators[5] are available, we demonstrate that it is possible to directly transfer the polarity information to other languages by using machine translation. Scandinavian languages and, more specifically, Danish (as Language A), Norwegian and Swedish (as Languages B and C) were chosen for these experiments as they are closely related and show a considerable degree of similarity. Algorithm 1 summarizes the process of automatically creating sentiment lexicons for the three Scandinavian languages.

The biggest challenge faced in order to map both resources is the fact that both SentiWordnet and DanNet contain several senses for each synset. It is therefore essential to disambiguate the senses in order to map both resources more precisely. In our approach, the information contained in the glosses or definitions of both resources, as the ones in (1) for the English wordsense *phenomenal* and (2) for the Danish wordsense *fænomenal*, are used for this purpose. Different measures are

[2] https://github.com/aesuli/sentiwordnet

[3] https://cst.ku.dk/english/projekter/dannet/

[4] For example, http://compling.hss.ntu.edu.sg/omw/

[5] In our work, we use Google Translation API (https://cloud.google.com/translate) as the translator.

Algorithm 1 Automatic Approach of Sentiment Lexicon Generation for Scandinavian Languages.

Input: English sentiment lexicon SL and target language dictionary D

Output: Sentiment lexicons $SWN_{1,2,3}$ in a target language family

Translate target language dictionary D into English D_{trans};
for *each lexical entry* w *in* D_{trans} **do**
  if *w.sense not in* SL **then** pass;
  else if *only one sense in* SL **then**
    $sentiment(w) \leftarrow SL(w).polarity\ weight$;
  else
    Compute $Similaritys(D_{trans}.gloss, SL.glosses)$ and get best match m;
    $sentiment(w) \leftarrow SL(m).polarity\ weight$;
  Copy $sentiment(w)$ to $D(w)$;
  $SWN_1 += D(w)$;
Translate target language sentiment lexicon SWN_1 into other languages $SWN_{2,3}$ within the same family;

Method	Syn	Sen
Jacquard	31,662	34,130
MTL (Subramanian et al., 2018)	33,257	35,724
Quick-thought (Logeswaran and Lee, 2018)	33,251	35,718
Skip-thought (Kiros et al., 2015)	33,257	35,724
SIF (Arora et al.,2017)	33,257	35,724

Table 3: Number of synsets and senses in the Danish sentiment lexicons created with the different methods used for mapping the glosses.

investigated to measure the similarity between glosses, as explained below.

(1) exceedingly or unbelievably great; "the bomb did fantastic damage"; "Samson is supposed to have had fantastic strength"; "phenomenal feats of memory"

(2) usædvanlig god, hurtig, påfaldende e.l.; ekstraord... (Brug: "Jeg er helt fænomenal til poker || meget tyder på, at den fænomenale vækst vil fortsætte i de kommende år")

Table 3 summarizes the results of the mapping through machine translation and different similarity measures (as explained in Section 5) between SentiWordNet and DanNet. Each row in this table specifies the number of synsets and senses. Approximately half of the number of synsets in DanNet is covered by using this approach. The numbers are very similar for most sentence similarity methods, except for Jacquard, which contains around 2,000 fewer synsets than the other lexicons. This is not surprising as Jacquard computes similarity on the basis of the tokens and not on the semantics, possibly discarding semantically similar sentences expressed using different lexicons.

5.1. Sentence similarity measures

5.1.1. Jaccard similarity

As a baseline to our approach, a simple Jacquard similarity measure was used to compare the glosses of those senses appearing more than once. Jaccard similarity is defined as the size of the intersection divided by the size of the union of two sets. Specifically, given two word sets A and B, the similarity between these two sentences can be derived as:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

The range is 0 to 1. The higher the score, the more similar the two sentences. This method is straightforward and simple as it is computed on the basis of word forms, but it cannot handle synonymy or capture the semantics of sentences.

5.1.2. Sentence embeddings

Besides Jaccard similarity, another well-known technique to measure sentence similarity is sentence embeddings. A variety of methods can be used to learn them. In this paper, we mainly focus on the following: simple averaging word embeddings, unsupervised approaches and multi-task learning approaches.

Bag-of-words approach directly averages a sentence's word embeddings. One of the strongest algorithms for generating semantic embeddings of sentences is smooth inverse frequency (SIF) proposed by (Arora et al., 2017). The main idea of this approach is to use pre-trained word embeddings such as GloVe[6]. This approach represents the sentence by a weighted average of the word vectors, and then performs a common component removal by removing the projection of the vectors on their first principal component. It has deeper and powerful theoretical motivations that rely on a generative model which uses a random walk on a discourse vector to generate text. Specifically, given the discourse vector c_s, the probability of a word w that is emitted in the sentence s is modeled by,

$$Pr\left[w \text{ emitted in sentence } s|c_s\right] = \alpha p(w) + (1-\alpha)\frac{exp(< \tilde{c}_s, v_w >)}{Z_{\tilde{c}_s}} \quad (2)$$

where $\tilde{c}_s = \beta c_0 + (1-\beta)c_s, c_0 \perp c_s$, $p(w)$ is the unigram probability of word, α and β are scalar hyperparameters, and $Z_{\tilde{c}_s} = \sum_{w \in \mathcal{V}} exp(< \tilde{c}_s, v_w >)$ is the normalizing constant (the partition function). This model has two types of "smoothing terms": i) an additive term $\alpha p(w)$ that allows words to occur even if their vectors have very low inner products with c_s. ii) A common discourse vector $c_0 \in \mathcal{R}^d$ serving as a correction term for the most frequent discourse that is often related to syntax. According to this model (2) the likelihood for the sentence is defined as

[6]It has 300-dimensional vectors that were trained on the 840 billion tokens from Common Crawl corpus and is publicly available at http://nlp.stanford.edu/projects/glove/.

$$p[s|c_s] = \prod_{w \in s} p[w|c_s] = \prod_{w \in s} [\alpha p(w) + (1-\alpha)\frac{exp(< \tilde{c}_s, v_w >)}{Z_{\tilde{c}_s}}] \quad (3)$$

Then, the sentence embedding is defined as the maximum likelihood estimate for the vector c_s that is generated, and it is updated by subtracting the projection of $\tilde{c}_s$'s to their first principal component.

Skip-thought vectors is an approach for learning unsupervised sentence embeddings proposed in (Kiros et al., 2015). It abstracts the skip-gram model to the sentence level. That is, rather than using a word to predict its surrounding context, it tries to encode a sentence to predict the sentences around it. The model consists of an RNN-based encoder-decoder trained to reconstruct the surrounding sentences from the current sentence. Specifically, given a sentence tuple (s_{i-1}, s_i, s_{i+1}), let $w_i^1, ..., w_i^N$ be the words in sentence s_i where N is the number of words in the sentence. Let $x_i^1, ..., x_i^N$ be the corresponding word embeddings. At each time step, the encoder produces a hidden state h_i^t which can be interpreted as the representation of the sequence $w_i^1, ..., w_i^t$. To encode a sentence, we iterate the following sequence of equations (dropping the subscript i):

$$\begin{aligned} \boldsymbol{r}^t &= \sigma(\boldsymbol{W}_r \boldsymbol{x}^t + \boldsymbol{U}_r \boldsymbol{h}^{t-1}) \\ \boldsymbol{z}^t &= \sigma(\boldsymbol{W}_z \boldsymbol{x}^t + \boldsymbol{U}_z \boldsymbol{h}^{t-1}) \\ \tilde{\boldsymbol{h}}^t &= tanh(\boldsymbol{W}\boldsymbol{x}^t + \boldsymbol{U}(\boldsymbol{r}^t \odot \boldsymbol{h}^{t-1})) \\ \boldsymbol{h}^t &= (1 - \boldsymbol{z}^t) \odot \boldsymbol{h}^{t-1} + \boldsymbol{z}^t \tilde{h}^t \end{aligned} \quad (4)$$

where $\tilde{h}^t$ is the proposed state update at time t, $\boldsymbol{z}^t$ is the update gate, $\boldsymbol{r}^t$ is the reset gate, $(\odot)$ denotes a component-wise product. Both update gates takes values between zero and one.

The decoder is a neural language model which conditions on the encoder output h_i. The computation is similar to that of the encoder except it introduces matrices C_z, C_r and C that are used to bias the update gate, reset gate and hidden state computation by the sentence vector. The decoder for the next sentence s_{i+1} involves the following sequence of equations (dropping the subscript $i+1$):

$$\begin{aligned} \boldsymbol{r}^t &= \sigma(\boldsymbol{W}_r^d \boldsymbol{x}^{t-1} + \boldsymbol{U}_r^d \boldsymbol{h}^{t-1} + \boldsymbol{C}_r \boldsymbol{h}_i) \\ \boldsymbol{z}^t &= \sigma(\boldsymbol{W}_z^d \boldsymbol{x}^{t-1} + \boldsymbol{U}_z^d \boldsymbol{h}^{t-1} + \boldsymbol{C}_z \boldsymbol{h}_i) \\ \tilde{\boldsymbol{h}}^t &= tanh(\boldsymbol{W}^d \boldsymbol{x}^{t-1} + \boldsymbol{U}^d(\boldsymbol{r}^t \odot \boldsymbol{h}^{t-1}) + \boldsymbol{C}\boldsymbol{h}_i) \\ \boldsymbol{h}_{i+1}^t &= (1 - \boldsymbol{z}^t) \odot \boldsymbol{h}^{t-1} + \boldsymbol{z}^t \tilde{h}^t \end{aligned} \quad (5)$$

Given $\boldsymbol{h}_{i+1}^t$, the probability of word w_{i+1}^t given the previous $t-1$ words and the encoder vector is

$$P(w_{i+1}^t | w_{i+1}^{<t}, \boldsymbol{h}_i) \propto exp(\boldsymbol{v}_{w_{i+1}^t} \boldsymbol{h}_{i+1}^t) \quad (6)$$

where $\boldsymbol{v}_{w_{i+1}^t}$ denotes the row of $\boldsymbol{V}$ corresponding to the word of w_{i+1}^t.

Thus, the objective optimized is the sum of the log-probabilities for the forward and backward sentences conditioned on the encoder representation:

$$\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, \boldsymbol{h}_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, \boldsymbol{h}_i) \quad (7)$$

The total objective is the above summed over all such training tuples.

Quick-thought vectors are a recent development of the Skip-thoughts vectors (Logeswaran and Lee, 2018). In this unsupervised sentence representation learning method, the task of predicting the next sentence given the previous one is reformulated as a classification task: the decoder is replaced by a classifier that has to choose the next sentence among a set of candidates. It can be interpreted as a discriminative approximation to the generation problem and achieves an order of magnitude speedup in training time.

Formally described, let f and g be parametrized functions that take a sentence as input and encode it into a fixed length vector. Let s be a given sentence. Let S_{ctxt} be the set of sentences appearing in the context of s (for a particular context size) in the training data. Let S_{cand} be the set of candidate sentences considered for a given context sentence $s_{ctxt} \in S_{ctxt}$. In other words, S_{cand} contains a valid context sentence s_{ctxt} (ground truth) and many other non-context sentences, and is used for the classification objective as described below. For a given sentence position in the context of s (e.g., the next sentence), the probability that a candidate sentence $s_{cand} \in S_{cand}$ is the correct sentence (i.e., appearing in the context of s) for that position is given by

$$P(s_{cand}|s, S_{cand}) = \frac{exp[c(f(s), g(s_{cand}))]}{\sum_{s' \in S_{cand}} exp[c(f(s), g(s'))]} \quad (8)$$

where c is a scoring function/classifier.

The training objective maximizes the probability of identifying the correct context sentences for each sentence in the training data D.

$$\sum_{s \in D} \sum_{s_{ctxt} \in S_{ctxt}} logP(s_{ctxt}|s, S_{cand}) \quad (9)$$

In our experiments, c is simply defined to be an inner product $c(u, v) = u^T v$. We use RNNs as f and g as they have been widely used in recent sentence representation learning methods. The words of the sentence are sequentially fed as input to the RNN and the final hidden state is interpreted as a representation of the sentence. We use gated recurrent units (GRU) as the RNN cell.

Multi-task learning approach can be seen as a generalization of diverse neural approaches to NLP tasks (such as skip-thoughts and machine translation) by combining the inductive biases of their training objectives in a single model. This approach builds representations that encode multiple aspects of the same sentence. In this paper, we adopt the MTL model proposed by (Subramanian et al., 2018) which leverages a one-to-many (a shared encoder and multiple task-specific decoders) multi-tasking learning framework to

learn universal sentence embeddings by switching between several tasks. The selected tasks (Skip-thoughts prediction of the next/previous sentence, neural machine translation, constituency parsing and natural language inference) share the same sentence embedding obtained by a bi-directional GRU.

Specifically, the input x and output y are sequences $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$. The encoder produces a fixed length vector representation $\boldsymbol{h}_x$ of the input, which the decoder then conditions on to generate an output. The decoder is auto-regressive and breaks down the joint probability of outputs into a product of conditional probabilities via the chain rule:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} P(y_i|y_{<i}, \boldsymbol{h}_x) \quad (10)$$

In this model, considering the computational speed, the encoder is a bidirectional GRU while the decoder is a unidirectional conditional GRU. During training procedure, every parameter is updated by uniformly sampling.

6. Evaluation

In order to evaluate these sentiwordnets, we performed a careful extrinsic rule-based evaluation. Firstly, all Danish resources, as presented before in Table 3, were evaluated on a dataset obtained from the Danish Tripadvisor. The Swedish and Norwegian resources, created by translating the best performing sentiment lexicon for Danish, were evaluated on similar datasets obtained from the Tripadvisor on these languages. To perform an additional evaluation on a genre of texts, the Norwegian resource was also evaluated on the Norwegian Review Corpus (NoReC) (Velldal et al., 2018). As a baseline for our experiments, we have also created plain sentiment lexicons by simply translating the English SentiWordNet into Danish, Norwegian, and Swedish without performing any kind of mapping.

6.1. Tripadvisor datasets

The Tripadvisor datasets were obtained by crawling reviews from restaurants in Copenhagen (for Danish)[7], Stockholm (for Swedish)[8], and Oslo (for Norwegian)[9]. In order to obtain a performance baseline of the original English SentiWordnet on the same sentiment analysis task, we also obtained an English Tripadvisor corpus including around 60k reviews and 1090 restaurants of London. Table 4 summarizes the Tripadvisor datasets used for evaluation. As can be seen from this table, a large number of reviews is included in these datasets, ranging from around 24 to 60 thousand reviews.

[7] https://www.tripadvisor.dk/Restaurants-g189541-Copenhagen_Zealand.html

[8] https://www.tripadvisor.se/Restaurants-g189852-Stockholm.html

[9] https://no.tripadvisor.com/Restaurants-g190479-Oslo_Eastern_Norway.html

Dataset	Restaurants	Reviews	Pos	Neg
Danish	2,045	44,260	41%	8%
Swedish	2,482	45,461	42%	8%
Norwegian	1,145	24,161	41.5%	8.5%
English	1,090	59,976	46.9%	3%

Table 4: Total number of restaurants and reviews in Tripadvisor (TA) datasets

The punctuation system used in Tripadvisor was converted to 1 if positive, and 0 if negative. In order to do that, only the scores representing a clear opinion were selected. Specifically, as Tripadvisor score system ranges from 1 to 5, the lowest number meaning the most negative opinion and the highest the most positive one, we considered 1 and 2 to exhibit negative sentiments, whereas 4 and 5 positive. Reviews with a score of 3 were left out from our datasets to avoid non-clearly opinionated reviews.

It is well known that the results of sentiment analysis can widely vary depending on the type or genre of text. For example, political texts are known for being difficult to classify automatically. In order to provide an evaluation on a different genre, the NoRec corpus was used for an additional evaluation of the Norwegian resource. This freely available dataset was created for the purpose of training and evaluating models for document-level sentiment analysis, and it contains 35,000 reviews collected from several of the major Norwegian news sources (Velldal et al., 2018)[10]. Following a Norwegian journalism convention, in this dataset the item under review is rated on a scale from 1 to 6. To perform the evaluation on this corpus following a similar criteria as in the Tripadvisor datasets, we transformed the scoring into a binary system, including 1 and 2 scores for positive and 5 and 6 for negative scores. Potentially not clearly opinionated reviews like 3 and 4 were left out from this dataset. As a result of this transformation, the total number of reviews used for evaluation is 17,512, among which 84.1% reviews are positive and 15.9% reviews are negative.

From the disproportionate ratio, we can see the sentiment polarity distribution is remarkably imbalanced across these two datasets, which will render the standard accuracy no longer reliable. There exists many ways to alleviate such phenomena, such as upsampling, downsampling, change training strategy and so on. In this paper, we adopt down-sampling of our datasets by randomly removing observations from the majority class and keeping the same number of observations with the minority class. The final performances are reported after 5 runs with the average test results.

6.2. Rule-based evaluation

A simple rule-based algorithm averaging the words with positive and negative scores in the sentences was

[10] https://github.com/ltgoslo/norec

Method	No lemmatization	Lemmatization
Jacquard	0.679	0.700
MTL	0.692	0.712
Quick-thought	0.699	**0.722**
Skip-thought	**0.671**	0.683
SIF	0.691	0.708

Table 5: F1-score of Danish sentiment lexicons obtained with the different sentence similarity methods on the Tripadvisor dataset.

used to evaluate these resources[11]. In order to filter out words by part of speech, several part-of-speech taggers and lemmatizers were considered in this evaluation. Precisely, Polyglot was used for Danish and Swedish morphological tagging (Al-Rfou et al., 2013) and Lemmy for Danish and Swedish lemmatization.[12] Experiments were carried out to assess the impact of lemmatization in the sentiment analysis tasks.

In contrast, to deal with the sentiment analysis of Norwegian, a part-of-speech tagger was trained exclusively for the purposes of this project. Specifically, the implementation of the Average Perceptron Tagger algorithm in NLTK was trained using the Norwegian UD treebank, which is a syntactic treebank of Norwegian.[13] This corpus contains around 300k words, 20k sentences, and it is manually annotated with morphosyntactic information of part-of-speech and syntactic categories. It includes text from several genres, blogs, parliamentary reports, and news from the main Norwegian newspapers (*Aftenposten*, *Dagbladet*, *Klassekampen*, *Sunnmørsposten*, and *VG*). The morphological tagset used here is mostly inspired by the Oslo-Bergen Tagger [14]. For the purposes of simplification and standardization, during training this tagset was transformed into a coarse-grained annotation standard that follows the UD scheme, including 17 POS tags[15]. After training, the part-of-speech tagger yields an accuracy of 96% on the test set from the Norwegian UD treebank. The tagger is also released together with sentiment lexicons under the MIT license.

Table 5 summarizes the results of the evaluation over all the Danish sentiment lexicons obtained by using different sentence similarity methods, on the Tripadvisor dataset. As can be seen from this table, the Danish sentiment lexicon obtained by using quick-thought vectors on the lemmas to compute the similarity between the glosses, yields the best F1-score[16] of 72.2%. In contrast, the worst performing sentiment lexicon, obtained with skip-thought vectors, yields 67.1%. Interestingly the Danish sentiment lexicon obtained by using the simplest sentence similarity measure – Jacquard method – that only checks the presence of similar words within the synset's definitions, yields 70%, only two points less than the best one.

Table 6 summarizes the evaluation results of the best performing Danish, Swedish and Norwegian sentiment lexicons. As can be seen from this table, the evaluation shows that the Norwegian sentiment lexicon performs quite well on the Tripadvisor dataset, with an F1-score of 72.5%, better than the Danish and Swedish lexicons. Conversely, the evaluation of the Swedish resource on the Swedish Tripadvisor corpus yields 71%, lower than the Danish sentiment lexicon. This might suggest that the translation from Danish to Swedish to render the Swedish resource is relatively worse. In contrast, the English SentiWordNet performs the best in the task of sentiment analysis. The difference of 3 points is, we believe, fair, given the limitations of NLP resources that could be used for pre-processing of Norwegian and Danish and the additional translation step. To compare to a simple translation approach, we created sentiment lexicons by directly translating SentiWordnet into Danish, Norwegian and Swedish. As can be seen from Table 6, the results show that our SWN methods perform better than translated sentiment lexicons (SL) in most cases on both F1-score, Precision and Recall. Compared with English SWN, we can see that SL methods in Danish, Norwegian and Swedish do preserve some accuracy to a certain extent, which verifies the main idea of this paper that a good translation mechanism can help, but our alignment mechanism based on sentence embeddings can further improve the performance without cumbersome manual annotation labor.

We also performed an evaluation of our method and three existing sentiment lexicons, Sentida (Lauridsen et al., 2019), NorSentLex (Barnes et al., 2019) and SenSALDO (Rouces et al., 2018b), in Danish, Norwegian and Swedish respectively. The results are shown in Figure 2. Our SWN method underperforms on Tripadvisor and NoRec datasets in Norwegian, which is reasonable considering the fact that our resource is created fully automatically without the help of human annotators compared with NorSentLex processed with fine human corrections. Besides, our method yields deteriorated F1-score, specifically 61.7% (micro average) and 58.3% (macro average), compared with the Tripadvisor dataset given the complexity of the texts included in the NoReC dataset. As it is well known the case that journalism is more challenging to classify into positive or negative categories. It is worth noting that, our SWN method outperforms Sentida and SenSALDO on Tripadvisor datasets even though the latter two also

[11] The code has been adapted from https://github.com/anelachan/sentimentanalysis/blob/master/sentiment.py.

[12] https://github.com/sorenlind/lemmy

[13] https://github.com/UniversalDependencies/UD_Norwegian-Bokmaal

[14] http://tekstlab.uio.no/obt-ny/

[15] universaldependencies.org/u/pos/index.html

[16] If not specified, F1-score represents the micro-averaged F1 score in this paper.

	F1-score		Precision		Recall	
	Micro avg	**Macro avg**	**Micro avg**	**Macro avg**	**Micro avg**	**Macro avg**
Danish SWN	**0.722**	0.709	**0.722**	**0.767**	**0.722**	0.721
Danish SL	0.722	**0.718**	0.722	0.737	0.722	**0.722**
Norwegian SWN	**0.725**	**0.717**	**0.725**	**0.751**	**0.725**	**0.725**
Norwegian SL	0.721	0.712	0.721	0.750	0.721	0.721
Swedish SWN	**0.710**	**0.705**	**0.710**	0.724	**0.710**	**0.709**
Swedish SL	0.701	0.688	0.701	**0.739**	0.701	0.701
English SWN	0.758	0.747	0.758	0.814	0.758	0.758

Table 6: F1-score, precision and recall of a rule-based classifier using the Danish, Swedish and Norwegian, and English sentiwordnets (SWN) and simply translated sentiment lexicons (SL) on the Tripadvisor datasets.

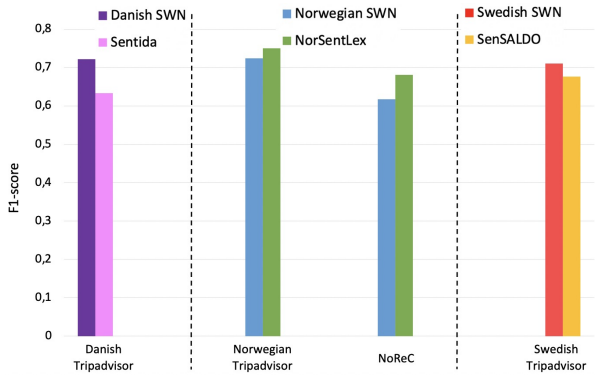

Figure 2: Comparison with the existing sentiment lexicons in Danish, Norwegian and Swedish.

include human annotations in the process of generating the sentiment lexicon. Furthermore, Sentida, NorSentLex and SenSALDO are purely sentiment lexicons with a list of the positive and negative words, whereas in the present paper our generated sentiwordnets keep other relevant information such as crosslinks to the English SentiWordNet and DanNet that might prove useful to NLP experiments on word sense disambiguation.

6.3. Error analysis

An analysis over a random selection of 30 reviews for each language and dataset shows that the most common source of error is due to cases where positive terms are used to express negative opinions. As noticed already by (Taboada et al., 2011), negative discourse tends to be expressed in euphemistic ways, which makes polarity more challenging to be identified in general. This blend of vocabulary might confuse the algorithm that just averages the number of positive and negative words, and the result turns out incorrect. An example in (1) from the Norwegian corpus illustrates this.

(1) Stilig restaurant men små porsjoner. Minimalistisk, kule lokaler på designhotellet Grims Grenka, Oslo sentrum. God mat, men den serveres i små små porsjoner. Gir inntrykk av at de er gjerrige og sparsommelige. Jeg går nok ikke tilbake hit.

'Stylish restaurant but small portions. Minimalist, cool premises at the design hotel Grims Grenka, central Oslo. Good food, but it is served in small portions. Gives the impression that they are stingy and thrifty. I'm probably not going back here.'

7. Discussion & Conclusion

This paper presents a simple but effective method to automatically create several sentiment lexicons on low resourced languages by using machine translation and sentence embeddings. The results of an extrinsic evaluation show close to state-of-the-art results, higher for those languages for which machine translators presumably perform best. A total number of four resources have been made freely available (MIT license) as a result of this project: a sentiment lexicon for Danish, for Swedish and for Norwegian, and a part-of-speech tagger for the Norwegian language. The method presented here can also be used for other under-resourced languages to create additional linguistic tools.

The principal disadvantage of our approach is the fact that it relies on the existence of a thesaurus in one of the target languages (Danish in our case), containing information about the senses and the definitions.However, the sentence embeddings methods suggest that maybe by considering the presence of these senses in a large corpus of the target language, and collecting the contexts or sentences (that would be like the definitions) in which they appear, similar results could be obtained.

8. Acknowledgements

This work is supported by the Research Council of Norway under Grant No. 245469 and 309834.

9. Bibliographical References

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

An, J., Kwak, H., and Ahn, Y.-Y. (2018). Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2450–2461.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, volume 10, pages 2200–2204.

Barnes, J. C., Touileb, S., Øvrelid, L., and Velldal., E. (2019). Lexicon information in neural sentiment analysis: a multi-task learning approach. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 175–186.

Elming, J., Plank, B., and Hovy., D. (2014). Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer.

Hamilton, W. L., Kevin Clark, J. L., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 595—-605. Austin, TX.

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. ACL.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373. Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Lauridsen, G. A., Dalsgaard, J. A., and Svendsen, L. K. B. (2019). Sentida: A new tool for sentiment analysis in danish. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(1):38–53.

Logeswaran, L. and Lee, H. (2018). An efficient framework for learning sentence representations. *International Conference on Learning Representations*.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 976–983.

Nusko, B., Tahmasebi, N., and Mogren, O. (2016). Building a sentiment lexicon for Swedish. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, July 11, 2016, Krakow, Poland*, number 126, pages 32–37. Linköping University Electronic Press.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a Wordnet for Danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Pedersen, B. S., Nimb, S., and Olsen, S. (2021). Dansk betydningsinventar i et datalingvistisk perspektiv. *Danske Studier*, (2021):72–106.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in Spanish. In *LREC*, volume 12, page 73.

Rosell, M. and Kann, V. (2010). Constructing a Swedish general purpose polarity lexicon random walks in the people's dictionary of synonyms. *Proceedings of Swedish language technology conference*, pages 19–20.

Rouces, J., Borin, L., Tahmasebi, N., and Eide, S. R. (2018a). Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. In *DHN*, pages 219–227.

Rouces, J., Nina Tahmasebi, L. B., and Eide, S. R. (2018b). Sensaldo: Creating a sentiment lexicon for swedish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. MIT press.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tsakalidis, A., Papadopoulos, S., Voskaki, R., Ioannidou, K., Boididou, C., Cristea, A. I., Liakata, M., and Kompatsiaris, Y. (2018). Building and evaluating resources for sentiment analysis in the greek language. *Language Resources and Evaluation*, 52(4):1021–1044.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Velldal, E., Øvrelid, L., Bergem, E. A., Stadsnes, C., Touileb, S., and Jørgensen, F. (2018). NoReC: The Norwegian Review Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the conference on empirical methods in natural language processing*, pages 553–561. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354.