Automatic recognition of German news focusing on future-directed beliefs and intentions

Judith Eckle-Kohler, Michael Kohler, Jens Mehnert

Darmstadt University of Technology, Department of Mathematics, Schloßgartenstrasse 7, 64289 Darmstadt, Germany

Abstract

We consider the classification of German news stories as either focusing on future-directed beliefs and intentions or lacking these. The method proposed in this article requires only a small set of labeled training data. Rather, we introduce German clues for the automatic identification of future-orientation which are used for automatic labeling of Reuters news stories. We describe the development of a high-precision procedure for automatic labeling in a bootstrapping fashion: A first version of the labeling procedure uses the absence of clues for future-directedness as indicator for non-future-directedness and is able to automatically label about one third of the Reuters news stories with high precision. Then a perceptron is applied to the automatically labeled news stories in order to semi-automatically acquire an additional set of clues for non-future-directedness. The second version of the labeling procedure additionally uses these clues and achieves remarkably improved results in terms of recall; it can even be extended by a guessing step to perform classification with an error of 22.5%. We also investigate another way to increase the recall by using the automatically labeled news stories as training data for statistical classifiers. Three different types of statistical classifiers are applied in order to address the question, which classifier is most suited for the text classification task considered. The best statistical classifier combined with the results of improved automatic labeling is able to recognize the two classes of news stories with an error of 19%.

Keywords: Text classification, automatic labeling, lexical-semantic verb classes, Boosting, neural networks.

Email addresses: ecklekohler@mathematik.tu-darmstadt.de (Judith Eckle-Kohler, Michael Kohler, Jens Mehnert), kohler@mathematik.tu-darmstadt.de (Judith Eckle-Kohler, Michael Kohler, Jens Mehnert), mehnert@mathematik.tu-darmstadt.de (Judith Eckle-Kohler, Michael Kohler, Jens Mehnert)

1. Introduction

In recent years the focus in automatic text analysis and classification has turned from pure topic classification to more fine-grained classification tasks with a special emphasis on non-factual text elements covering subjectivity identification [44], opinion detection [47], [10], [14], semantic polarity analysis [37], [23], [38], [12], certainty identification [29], and perspective identification [21].

While previous work has investigated the automatic recognition and analysis of a subset of non-factual text elements, namely opinions, evaluations and sentiments, the work presented in this article focuses on future-directed beliefs (like expectations, speculations, forecasts) and intentions (like plans, promises, suggestions) which belong to the category of non-factual text elements as well, and whose automatic identification has largely been neglected so far.

In this article, we concentrate on the automatic classification of future-directedness on document level, using the new Reuters corpus (Reuters Corpus 2000) which consists of mostly short news stories.

Following are examples of future-directed news stories from the English Reuters corpus. The first story is an example of a news story focusing on future-directed beliefs:

(1) French chamber of commerce sees 2.2 pct GDP growth.
PARIS 1997-04-01
The Paris Chamber of Commerce predicted on Thursday that French gross domestic product would grow 2.2 percent this year but it did not expect France to meet the Maastricht target for monetary union limiting the government deficit to three percent of GDP. The forecast from the Chamber's economic forecasting branch COE is below the government's 2.3 percent GDP growth forecast and a 2.5 percent forecast issued on Thursday by another private forecaster, the OFCE. The COE said the recent rise in the dollar and lower interest rates were the main factors behind the acceleration in GDP growth. France's GDP grew 1.3 percent in 1996. It said tight public spending, reflecting the government's effort to meet the Maastricht criteria, would have a dampening impact on growth. But it said it expected the deficit to total 3.3 percent of GDP. The COE forecast 2.6 percent GDP growth for 1998. [25]

The next example is a news story about future-directed intentions:

(2) Japan to set fuel efficiency goals for diesel cars.
TOKYO 1997-04-01
A Japanese cabinet meeting on Tuesday approved a plan to set fuel efficiency goals for vehicles with diesel engines, as part of a comprehensive energy conservation package, government officials said. Under the plan, fuel efficiency targets will be set for passenger cars

with diesel engines by fiscal 1998/99, beginning April 1998, and for diesel trucks by fiscal 2000/01, they said. The targets would make Japan the first country with energy efficiency goals for diesel cars, officials at the trade and transport ministries said. Details of the new regulations, including how soon and by how much fuel efficiency should be improved, will be worked out at a joint panel of the two ministries. Japan has already set fuel efficiency targets for passenger vehicles and trucks with gasoline engines. [25]

In contrast, following is an example of a non-future-directed, factive news story reporting an event:

(3) Power outages hit 100,000 in N.Y., 10 pct of Mass.
POUGHKEEPSIE, N.Y. 1997-04-01
Orange; Rockland Utilties said Tuesday the blizzard that swept across the Northeast overnight left some 100,000 customers without power in the mid-Hudson Valley region. "We have restored service to about half of those customers," Mike Donovan, a spokesman for the utility told Reuters in a telephone interview Tuesday afternoon. Massachusetts Gov. William Weld, who declared a state of emergency in the Commonwealth, told reporters that about 10 percent of the state's utility customers were without power. "It's not just the distribution lines that are down," Weld told reporters at a news conference in Framingham, Mass. "Some transmission lines are down as well and that's why you have as many as a quarter of a million households without electricity." [25]

The automatic identification of future-directedness on document level might be useful for text filtering which could be applied, for example, to news stories or to Weblogs: texts identified as predominantly future-directed are not suited as a basis for extracting facts, but they are instead a good source for retrieving information about company strategies and expectations, analyst speculations, and political plans, or, in the case of Weblogs, peoples' intentions, expectations and future plans (see [7]). The automatic classification of Reuters news stories as future-directed or not future-directed as described in this article yields a large number of classified news stories. These labeled documents can be used as a basis for developing methods for classification of future-directedness on sentence level. A sentence level classification of future-directedness helps to separate factual content from non-factual content, which can be used to support many natural language processing tasks, such as information extraction, summarization, question answering, and semantic tagging.

Our approach to the automatic identification of future-directed beliefs and intentions is based on the use of lexical clues for future-directedness. A major part of these clues constitute lexical-semantic classes, capturing the close relationship between the syntax and semantics of verbs, nouns, and adjectives. Future-directedness in German often is indicated by classes of verbs, nouns, and

adjectives, sharing common behaviour with respect to possible alternations of several types of clausal complements: all of them can occur with to-infinitives, while other types of clausal complements are restricted or even not possible.

Using mainly lexical-semantic classes as clues for a specific text classification task has two major advantages. Firstly, lexical-semantic classes capturing the syntax-semantics relationship have already been contructed in several languages, varying in size and coverage, though: for English there is Levin's large verb classification [19] and its extension by [17] based among other sources on Rudanko's work on classes of verbs taking clausal complements, see [30], [31]. There are also verb classifications for Spanish [40] and French [32], and for German, see [36] and [4]. Secondly, and even more important, the definition of lexical-semantic classes by purely syntactic properties makes it possible to automatically acquire these classes from corpus data, which has been shown to be feasible by several attempts recently made in this area, see [22], [35].

The rest of this paper is structured as follows: The manual annotation of data used for our classification experiments is described in section 2. Section 3 introduces German clues for future-directed beliefs and intentions which constitute lexical semantic classes. The high-precision algorithm for the identification of future-directed and non-future-directed news stories that is based on the presence vs. absence of clues for future-directedness is presented in section 4. In section 5 we use the automatically labeled news stories as training data for three kinds of classifiers, in order to improve the recall: nearest-neighbor classifiers, neural network classifiers, and decision stumps trained by *AdaBoost*. Section 6 describes major improvements of the automatic labeling algorithm based on the acquisition of new clues: lexical clues for non-future-directedness are acquired semi-automatically by applying a perceptron to the automatically labeled data already available. With these additional clues, the improved labeling algorithm can be extended to perform classification by adding a guessing step, resulting in an error rate which is approximately equal to that of the statistical classifiers. With the improved automatically labeled training data, classification with statistical classifiers now yields the best results, when the statistical classifiers are applied only to texts not automatically labeled yet. Section 7 discusses related work and describes possible extensions and future work.

2. Manual classification of news

We conducted our experiments using the German part of the new Reuters corpus [26]. The German part consists of 116,212 manually categorized newswire stories produced by Reuters journalists between August 26, 1996 and August 19, 1997. The stories have an average length of 215 tokens.

A large hierarchical category taxonomy has been used to manually assign category codes for topic, industry and region. The topic code hierarchy includes two sub-categories named 'COMMENT/FORECASTS' and 'STRATEGY/PLANS' which are assigned, among others, to future-directed stories belonging to the parent category 'CORPORATE/INDUSTRIAL'. Unfortunately, we could not use these two categories for evaluation purposes, because they are also attached

to stories without future-directed content. Therefore we had to manually classify a set of news stories in order to obtain a test set to be used as gold standard for evaluation. We have developed the following annotation guidelines for manual classification:

1. Category F is to be assigned, if the story focuses on future-directed speculations, forecasts, expectations, intentions, plans, opinions on plans, negotiations about plans, agreement on plans, decisions. Background facts may also be given, as well as reasons or statements, but these are not the focus of the story.
2. Category N is to be assigned, if
   (a) the story focuses on factual information, such as reports on events or company reports. Again, expectations or future plans may be included, but these are not the focus of the story.
   (b) the story focuses on statements, reactions, evaluations, assessments, possibly including (but not focusing on) expectations and future-directed plans.

Since the Reuters news stories are rarely homogeneous with respect to future-directedness, the guidelines make use of the notion *focus* of a news story. Whether the focus of a given news story was future-directed or not, was the most difficult part of the manual annotation task and a source of disagreement between the annotators, too. However, the annotators shared the intuition, that often the focused content of a news story was presented first, whereas the last part contained background material. This property of journalistic writing style has been described before in the literature on summarization, cf. [2], [20].

We started with a training phase, where three annotators (**E**, **K**, and **M**) acquired a common understanding of the annotation guidelines by first annotating a set of 200 stories randomly chosen from the Reuters corpus and then discussing the evaluation results. For the subsequent annotation phase, we used a fresh set of 400 randomly selected news stories. This set was divided in three subsets of 200, 100, and 100 news stories in order to reduce the annotation effort. The first subset of 200 news stories was annotated by all three annotators independently, while the smaller subsets were annotated only by two annotators independently. Pairwise annotation agreement for the subset of 200 news stories was 86% (**E**, **K**), 88% (**E**, **M**), and 84% (**K**, **M**). The second subset of 100 stories was annotated by annotators **E** and **M** independently, resulting in 81% annotation agreement. Annotators **K** and **M** annotated the third subset of 100 news stories with annotation agreement of 80%. The resulting average pairwise agreement is 83.8%. The calculation of Cohen's Kappa for annotators **E** and **K**, **E** and **M**, **K** and **M** [1] yielded the values 0.71, 0.69, and 0.62, respectively. This indicates that the investigated classification task on document level is difficult, but feasible by humans, given the above guidelines. Difficulties the annotators had in making their annotation decision were mainly due to the fact that the news stories were rarely homogeneous, often containing both future-directed and non-future-directed phrases and sentences at the same time.

Based on the manual annotation, the gold standard was constructed as follows:

- for the first 200 stories, the majority annotation decision went into the gold standard;
- for the second and third subset of 100 stories, the remaining third annotator labeled those stories, where the two annotators had disagreed. Again, the majority annotation decision went into the gold standard.

The resulting gold standard consisting of 400 news stories contains 41% future-directed stories.

3. German lexical-semantic classes indicating future-directedness

In this section, we introduce German lexical-semantic classes of verbs with future-directed meaning. First we list syntactic properties which the verbs belonging to these classes have in common. Those common syntactic properties provide the basis for the semi-automatic acquisition of the verb classes from corpus data. Then we describe the process of manually filtering the lexical-semantic classes in order to get lexical clues for future-directedness. Finally we present a taxonomy of the resulting lexical-semantic classes of future-directed verbs.

3.1. Syntactic alternations

Among the verbs taking to-infinitives in German, a high number of verbs with future-directed meaning can be identified. Consider, for example, the verbs *beabsichtigen* 'intend' and *hoffen* 'hope':

> *Er beabsichtigt, das Buch morgen zu kaufen.* 'He intends to buy the book tomorrow.'
>
> *Er hofft, das Buch morgen zu bekommen.* 'He hopes to receive the book tomorrow.'

Most of the verbs taking to-infinitives in German can alternatively take certain types of finite clauses as complements, for example that-clauses or dependent declarative clauses. Consider again the verbs *beabsichtigen* and *hoffen*:

> *Er beabsichtigt, dass er das Buch morgen kauft.* 'He intends that he will buy the book tomorrow.'
>
> *Er hofft, dass er das Buch morgen bekommt.* 'He hopes that he will receive the book tomorrow.'
> *Er hofft, er bekommt das Buch bald.* 'He hopes he will receive the book soon.'

Other kinds of finite clauses occurring as complements in German are wh-clauses and whether-clauses. Both *beabsichtigen* and *hoffen* are not able to take either of these clauses, as the following examples show:

- ⋆ *Er beabsichtigt, wo er das Buch kauft.* 'He intends where he will buy the book.'
- ⋆ *Er beabsichtigt nicht, ob er das Buch kauft.* 'He does not intend, if he will buy the book.'
- ⋆ *Er hofft, wann er das Buch bekommt.* 'He hopes, when he will receive the book.'
- ⋆ *Er hofft, ob er das Buch bekommt.* 'He hopes, if he will receive the book.'

This is due to the fact, that wh-nominal complements are typical of factive verbs [42], whereas future-directed verbs clearly belong to the group of the non-factives.

However, there are many verbs taking to-infinitives that do not convey future-directed meaning. They typically allow those to-infinitives in past tense, that are used to express something clearly past-related, as indicated, for example, by the use of the adverb *yesterday*. Examples of such verbs are *bereuen* 'regret' and *behaupten* 'claim':

- *Er bereut (behauptet), das Buch gestern gekauft zu haben.* 'He regrets (claims) having bought the book yesterday.'

In contrast, future-directed verbs are not able to take to-infinitives in past tense used to express something past-related, as opposed to past-in-the-future (time which is in the past when seen from a viewpoint in the future). Consider the following examples:

- ⋆ *Er plant, das Buch gestern gekauft zu haben.* 'He plans having bought the book yesterday.'
- *Er plant, das Buch nchste Woche schon gekauft zu haben.* 'He plans having already bought the book next week.'

To summarize, at least six different types of realizations of a clausal complement can be distinguished in German:

1. to-infinitive in present tense
2. to-infinitive in past tense
3. that-clause
4. dependent declarative clause
5. wh-clause
6. whether-clause

Table 1: Example: alternation behaviour of *beabsichtigen* and *hoffen*

	to-inf. present	to-inf. past	that-clause	dep. decl. clause	wh-clause	whether-clause
beabsichtigen	Y	N	Y	N	N	N
hoffen	Y	N	Y	Y	N	N

The alternation behaviour of the verbs *beabsichtigen* and *hoffen* with respect to these six syntactic alternations can be characterized by Table 1, where possible realizations of the clausal complement are marked as **Y** (Yes), and realizations not possible as **N** (No).

If verbs taking clausal complements were to be classified according to their syntactic alternation behaviour with respect to these six clause types, it would theoretically be possible to end up with 64 different classes of verbs. But due to the different meanings the different clause types convey, far less classes occur in reality. [4] has manually examined the alternation behaviour of 784 verbs taking clausal complements with respect to the six clause types introduced above. These 784 verbs come from a German broad coverage subcategorization lexicon which has been semi-automatically acquired from a German newspaper corpus consisting of 200 million words[1]. [4] lists only 25 syntactic verb classes among which are 6 classes of verbs taking to infinitives in present tense but not in past tense and not allowing wh-nominals.

As [4] has argued, the verbs in each syntactic alternation class share a number of meaning components and therefore constitute lexical-semantic classes in Levin's sense [19], that is, their alternation behaviour with respect to possible alternations of the six types of clausal complements is largely determined by their meaning.

3.2. Criteria for manual filtering

By taking a closer look at the lexical-semantic verb classes listed in [4], we found that most of the verbs from the six classes allowing to-infinitives in present tense, but not in past tense and not allowing wh-nominals, indeed have future-directed meaning. We manually selected the verbs to be used as clues for future-directedness in news stories. It was necessary to exclude a small number of frequently occurring polysemic verbs like, for example, the verb *fragen* 'ask', which can have the meaning of 'ask a question', but which can also be used in the sense of *bitten* 'ask somebody to do something'. Other polysemic verbs could nevertheless be used as clues, because they have some predominating sense in news stories.

[1]The acquisition methods described in [4] put a special focus on high accuracy and fine-grained distinctions of subcategorization frames: the lexicon contains subcategorization frames of verbs, nouns, and adjectives and distinguishes between syntactic type and syntactic function of the complements, differentiating not only between the various possible types of prepositions, but also between 6 clause types, as well as correlates and reflexives.

However, we also found verbs with future-directed meaning in the classes of verbs taking wh-nominals (but not allowing to-infinitives in past tense), in spite of the fact, that the ability of a verb to take wh-nominal complements indicates factivity. Therefore we inspected those verb classes in order to identify further future-directed verbs. As test for future-directedness we checked, whether the adverb *tomorrow* could be used in the complement clause. Consider, for example, the verb *entscheiden* 'decide':

> *Er entscheidet jetzt schon, wer* **morgen** *den Preis bekommt.* 'He is deciding who will get the prize tomorrow.'

The manual inspection procedure resulted in 9 syntactic alternation classes of verbs with future-directed meaning.

3.3. Taxonomy of future-directed verbs

We present a taxonomy of the resulting lexical semantic classes in order to provide a characterization of the verbs belonging to these classes. Two main classes are distinguished: firstly, verbs marking actions as future-directed and secondly, verbs marking propositions as future-directed. This seems to be a natural distinction, because there are modal auxiliaries in German which express similar meaning: on the one hand the modal auxiliaries *wollen* 'will, want', *sollen* 'should', *müssen* 'must', which can as well be used to mark actions as future-directed, and on the other hand particular forms of *können* 'can, could' and *dürfen* 'may, might', namely *könnte(n)* and *dürfte(n)*; they can as well be used to mark propositions as future-directed.

The distiction between actions and propositions in this context has also been proposed by [11] who distinguishes between actional attitude verbs (e.g. *beabsichtigen* 'intend') and propositional attitude verbs (e.g. *hoffen* 'hope'). Tables 2 and 3 show the alternation behaviour of the resulting lexical-semantic classes of future-directed verbs. The complete lists of verbs are included in appendix A.

Most of the verbs in the lexical semantic classes are either attitude verbs, for example, *beabsichtigen* 'intend', *hoffen* 'hope', or speech act verbs, for example, *versprechen* 'promise', *vorhersagen* 'predict'. These two types of verbs are related, because "speech is essentially the expression of thought" [41]. Only the verbs in the *consequence*-class are different, because they are able to connect two propositions, marking one as the (future-directed) consequence of the other, for example, *hinauslaufen auf* 'amount to something', *führen zu* 'result in something'.

4. Automatic labeling of news using clues for future-directedness

In this section, we describe the development of an algorithm for automatic labeling of a subset of a given set of news stories as future-directed or non-future-directed. We evaluate the resulting algorithm on the gold standard using precision and recall measures. For the development of the automatic labeling

Table 2: German verbs marking actions as future-directed

	to-inf. present	to-inf. past	that-clause	dep. decl. clause	wh-clause	whether-clause
refuse-class	Y	N	N	N	N	N
consider-class	Y	N	N	N	N	Y
hesitate-class	Y	N	N	N	Y	Y
intend-class	Y	N	Y	N	N	N
try-class	Y	N	Y	N	N	Y
promise/urge-class	Y	N	Y	Y	N	N
suggest-class	Y	N	Y	Y	N	Y
plan/agree-class	Y	N	Y	N	Y	Y
decide-class	Y	N	Y	Y	Y	Y

Table 3: German verbs marking propositions as future-directed

	to-inf. present	to-inf. past	that-clause	dep. decl. clause	wh-clause	whether-clause
consequence-class	Y	N	Y	N	N	N
hope-class	Y	N	Y	Y	N	N
wait-class	Y	N	Y	N	Y	Y
predict-class	Y	N	Y	Y	Y	Y

algorithm we used a training set of 200 news stories which are different from the gold standard and which have been annotated by one annotator alone.

The algorithm for automatically identifying news stories of category F (future-directed) and N (non-future-directed) has been designed to achieve maximum precision, whereas low recall has been tolerated. Such a high-precision algorithm can be used to automatically label a subset of the Reuters stories, thus automatically creating training data for statistical classifiers. The core of the labeling algorithm is a simple decision procedure that counts the number of clues for future-directedness in a given document.

4.1. Using only verbs as lexical clues

Relying on the observation of [15] that verbs play an important role in the automatic identification of document types, we started our experiments with the classes of future-directed verbs introduced in section 3 as clues for future-directed beliefs and intentions in news stories.

The main task of the automatic labeling algorithm is to count the number of the selected verb lemmas in a given news story. In order to identify verb lemmas in German text, a certain amount of morpho-syntactic preprocessing of the text is required, involving at least Part-Of-Speech-tagging and lemmatization. For both tasks we chose Schmid's freely available decision-tree-tagger [34][2], which

[2]Downloadable at http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html

uses the STTS-Tagset (see [33]) consisting of 54 tags. A third preprocessing step is required due to the fact that in German there is a high number of verbs with separable prefix. In verb-second sentences these prefixes are separated from the verb stem. Consider, for example, the verb *vorhaben* 'intend': *Er* **hat** *immer noch* **vor**, *das Buch zu kaufen.* 'He still intends to buy the book.' In our implementation, the identification of verbs with separated prefixes relies on finite-state patterns over Part-Of-Speech-tags.

The algorithm counts all occurrences of the selected verbs as clues for future-directedness, regardless of their tense and voice. This is a simplification, because not all occurrences of the verb clues do actually express something future-related with respect to the present point of view. Consider the following examples:

> *Er hatte erwartet, zur Party eingeladen zu werden.* 'He had expected to be invited to the party.'
>
> *Er hatte vor, das Buch zu kaufen.* 'He intended to buy the book.'

Here, the to-infinitive-taking verb is in past tense and hence is used to express future-in-the-past, that is, time which is in the future when seen from a viewpoint in the past. But since tense annotations have not been easily available, we decided to count the number of all forms of the future-directed verbs in order to support the identification of the non-future-directed news stories which is based on the absence of clues.

We added a weight $k > 1$ to those verb clues that occur in the first quarter of a news story. This weight models the observation that in journalistic writing style the focus of a news story is typically presented at the beginning. Then the numbers of clues are added up and the sum is divided by the length of the story, yielding a relative clue frequency c for each story. Finally c is compared to a threshold parameter $t > 0$ in order to determine a label:

$$\text{if } c > t \text{ then } class := F$$

$$\text{elsif } c = 0 \text{ then } class := N$$

$$\text{else } class :=< unknown >$$

We used the training set of 200 news stories to determine the best values for the parameters k (weight of verb clues in the first quarter) and t (threshold) where precision for class F and N was maximal[3]. The parameter values $k = 4$ and $t = 0.05$ achieved best precision on the training set of 200 news stories for both classes. Using these parameter values, the labeling algorithm has been evaluated against the gold standard. Table 4 shows the resulting precision and recall.

It is clear that the precision for non-future-directed news stories is not high enough, having in mind to apply the algorithm for automatic labeling.

[3]Precision of class F is calculated as the percentage of class F labels that the algorithm proposes which are correct, recall of class F is calculated as the percentage of class F labels that the algorithm proposes. Precision and recall of class N are computed analogously.

Table 4: Using lexical-semantic classes of verbs as clues

evaluation measure	results for threshold $t = 0.05$
class F, precision	100%
class F, recall	1,83%
class N, precision	71, 49%
class N, recall	66,95%

4.2. Extending the clue list

Using just verbs as clues, we got only moderate precision for the class of non-future-directed documents. This is due to the fact that the labeling algorithm bases its decision for class N on the absence of future-directed clues. Obviously, the list of verb clues is incomplete, since there are, among others, many deverbal, future-directed nouns, taking to-infinitives as complement, too, for example, *Absicht* 'intention' or *Hoffnung* 'hope'. Nouns of this sort occur frequently in news stories, since they are written in a noun-based style.

Therefore we extended our clue list by adding future-directed nouns and adjectives to it, which are mainly derived forms of the verb clues, see complete clue list in appendix A.

Because the resulting precision for class N was still not high enough, we subsequently added two other groups of lexical items to the clue set in order to get a further reduction of the number of news stories that are erroneously identified as non-future-directed. Firstly, we used present tense forms of the modal auxiliaries *wollen* 'will, want', *sollen* 'should', *mssen* 'must', as well as certain forms of *knnen* 'can, could' and *drfen* 'may, might', namely *knnte(n)* and *drfte(n)*, see list in appendix B. *Wollen*, *sollen* and *mssen* in German can be used to express future-directed intentions as in *Er will nchstes Jahr wieder kandidieren.* 'He wants to candidate again next year.' The modal forms *knnte(n)* and *drfte(n)* have the meaning 'possibility' and can therefore be used to express future-directed beliefs, for example:

> Die Aktie knnte morgen wieder steigen. 'The share could be raising tomorrow again.'
>
> *Der Bau der Pipeline drfte zgig vorangehen.* 'The construction of the pipeline might proceed quickly.'

Secondly, we added a number of temporal adverbs and expressions which we have frequently observed in the training data of 200 news stories. Examples are *knftig* 'in future', *frhestens* 'at the earliest', *sptestens* 'at the latest', see complete list in appendix C. The results of using all three types of clues in concert are shown in Table 5.

With this final version of the clue list, the labeling algorithm achieves a precision of 91% for class N which is a significant improvement compared with the first experiment where just verbs were used as clues. At the same time the

Table 5: Final version: Using lexical-semantic classes, modals and temporal expressions as clues

evaluation measure	results for threshold $t = 0.07$
class F, precision	96.67%
class F, recall	17.68%
class F, F-measure	29.89%
class N, precision	91.36%
class N, recall	31.36%
class N, F-measure	46.69%

recall for class F is now about 15% higher than in our first experiment, while the precision is still at 96%[4].

The application of the labeling algorithm to all 116,212 news stories from the Reuters corpus yielded a set of 37,703 labeled documents (corresponding to 32,44%). This set of labeled documents contains 31% future-directed ones. As the recall of the algorithm for class F is lower than that for class N, the distribution of the N- and F-labeled documents is biased towards the N-labeled documents, compared with the Gold Standard.

4.3. Evaluation

Our experiments with different clue sets have shown, that the high precision value for class N is the result of combining three types of lexical clues: predicates taking to-infinitives, as well as modal auxiliaries and temporal adverbials. This experimental result shows that future-directedness in German is expressed via a combination of three types of future-directed clues: lexical-semantic classes of verbs, noun, and adjectives being able to take to-infinitives as complement, as well as temporal adverbials and specific forms of modal auxiliaries.

The approach of using the absence of clues for future-directedness as indicator for non-future-directedness is problematic, because the precision for labeling class N documents entirely depends on the completeness of the list of future clues. We expect that this problem will become worse when the algorithm is applied to other text types. Section 6 describes a solution to this problem.

5. Application of statistical classifiers

The algorithm described in the previous section achieves very high precision, but it leaves a large part of the news stories unlabeled. In order to improve the recall, we combine the automatic labeling algorithm with statistical classifiers. The basic idea is to automatically create labeled training data for a statistical

[4]We also tested a version of the clue set consisting only of the above mentioned modal verbs. But with this small clue set, there was a decrease in precision of about 10% for class N (compared with Table 5).

classifier which is then used to classify **all** news stories as either future-directed or non-future-directed. As pointed out by a referee, alternative methods which could have been applied here are semi-supervised learning techniques, using both labeled and unlabeled data points for training, see e.g. [48] for an overview.

It is well-known in non-parametric statistical classification that there is no single best classifier which is superior to all other classifiers in all situations, cf. Problem 7.3 in [3]. Therefore we examine the performance of three different statistical classifiers, namely nearest-neighbor classifiers, neural network classifiers, and decision stumps trained by *AdaBoost*, all of which are described in detail below.

We use the 37,703 automatically labeled Reuters documents to create training data for the statistical classifiers. Each document is represented by a $d = 500$-dimensional real vector containing relative frequencies of the first d most frequent token types, including words and punctuation marks, but excluding numbers. All uppercase characters have been converted to lower case, because the Reuters news stories are rather short (215 tokens on average). We do not, however, use lemmatization or a stoplist in order to preserve the information contained in the various word forms. Thus we get a training sample consisting of $n = 37,703$ data points, where each data point has as x-component a d-dimensional real vector, and as y-component a label from $\{0, 1\}$. Here 0 is the label for class F and 1 is the label for class N.

Two statistical classifiers have been applied to the training set of all 37,703 data points, namely nearest-neighbor classifiers and the *AdaBoost* algorithm. Firstly, we consider the classical nearest-neighbor classification method, which has as parameter a natural number k between 1 and n. It determines the class of a d-dimensional real vector z by choosing those k data points from the training data whose x-values are closest (with respect to the Euclidean metric) to the given z, and by assigning that class to z occurring most often among these k data points. In order to choose the parameter k in a data-dependent way, we use splitting of the sample, that is, we split our data in two parts of size $18,852$ and $18,851$, generate for $k \in \{1, 2, 4, 8, 16, 32\}$ the corresponding nearest-neighbor classifiers based on the training data, and finally choose the classifier that has the smallest empirical error on the second part of the data.

Secondly, we use the *AdaBoost* algorithm, cf. [5], to fit a linear combination of decision stumps to the data. A short description of the *AdaBoost* algorithm is given in Figure 1.

The simple classifiers we use here are decision stumps, which are decision trees with only one inner knot. At this inner knot the d-dimensional real vector space is divided in two regions, and on each of these two regions the class occurring among the training data in this region most often is used as prediction. The division of the d-dimensional vector space is done by choosing one of the d-components of x and by splitting this component at one splitting point into two parts. Here we chose the number of the component and the location of the splitting point by minimization of the empirical error on the training data. This is in contrast to standard decision trees, where usually some kind of entropy is minimized in order to determine the number of the component and the splitting

1. Initialize the observation weights: $w_i = 1/n$ $(i = 1, \ldots, n)$
2. For $k = 1$ to K:
   (a) Fit a simple classifier g_k to the training data using weight w_i for the i-th data point.
   (b) Compute
   $$err_k = \frac{\sum_{i=1}^{n} w_i \cdot I_{\{y_i \neq g_k(x_i)\}}}{\sum_{i=1}^{n} w_i}.$$
   (c) Set
   $$\alpha_k = \log \frac{1 - err_k}{err_k}$$
   and update the weights by
   $$w_i \leftarrow w_i \cdot \exp\left(\alpha_k \cdot I_{\{y_i \neq g_k(x_i)\}}\right)$$
   $(i = 1, \ldots, n)$.
3. Output the classification rule
$$g(x) = \begin{cases} 1 & if \sum_{k=1}^{K} \alpha_k \cdot g_k(x) \geq 1/2, \\ 0 & else. \end{cases}$$

Figure 1: The *AdaBoost* algorithm.

point. However, it is more appropriate here, because a stepwise minimization of the empirical error is not necessary for a single decision stump.

Finally, we fit a feed-forward neural network with one hidden layer of the form
$$f(z) = c_0 + \sum_{k=1}^{K} c_k \cdot \sigma\left(a_k \cdot z + b_k\right)$$
(with σ chosen as logistic sigmoid function $\sigma(u) = 1/(1+exp(-u))$, real numbers b_k, c_k, d-dimensional real vectors a_k and a natural number K) to the data to estimate the a posteriori probability of a d-dimensional real vector z. Since this a posteriori probability can be considered as a regression function (cf., e.g., section 1.4 in [8]), we determine the coefficients a_k, b_k and c_k by minimizing the empirical L_2-risk of the resulting function via the backpropagation algorithm (cf., e.g., section 11.4 in [9]). Here the parameter $K \in \{1, \ldots, 5\}$ of the estimate is again determined by splitting of the sample. The resulting classifier predicts class 1, if the estimated a posteriori probability of this class is greater than 0.5, and class 0 otherwise. From the 37,703 automatically labeled Reuters documents we randomly chose n=10,000 as training data for the neural network in order to limit training time.

We have implemented the three classifiers in the *C++* programming language. More precisely our test system is an *AMD* 64 bit *CPU* (2000 MHz) and 1 GByte RAM, running with *Gentoo Linux OS* (*Kernel 2.6.16*). As *C++* com-

piler we use *GCC 3.4.5* with compiler flags *"march=athlon64 -O6 -DNDEBUG"*. With the training parameters given above, k nearest neighbor is running 1 minute and 53 seconds until termination, consuming most of the time by reading the input data. The running time of *AdaBoost* is linear in the number of decision stumps. With 100 stumps we get a running time of 40 minutes. As for our neural network implementation, we limited the number of backpropagation steps (one backpropagation step recomputes the neural network weights) to 50,000 and got a running time of 24 minutes for our best neural network with one hidden neuron.

Table 6 gives the results of applying the three classifiers to our training data (column **SC**). Again, the classifiers are evaluated against the gold standard. We have also evaluated a combination of the results of automatic labeling and the statistical classifiers (column **AL-SC**), i.e., choosing the label of the statistical classifier only in cases, when there was no automatically assigned label available.

The statistical classifiers are able to recognize the two classes of news stories with significantly higher recall, but still acceptable precision, compared with automatic labeling. If we compare the results of the k nearest neighbor with the two modern classifiers *AdaBoost* and the neural network, then k nearest neighbor performs worst, while both *AdaBoost* and the neural network show significantly better performance. The results for *AdaBoost* and the neural network are equally good with regard to the error measure. The combination of the statistical classifiers with the automatically assigned labels yielded a remarkable improvement for k nearest neighbor, while the results for *AdaBoost* and the neural network did not improve. Surprisingly the combination of the worst statistical classifier (i.e., k nearest neighbor estimate) with the automatic labeling is better than all other estimates. A possible explanation is that there is a systematic difference between the training data and the rest of the data without automatically attached label, which is due to specific labeling technique based on clue words. While the modern statistical classifiers are very well adapted to the automatically created training data, an estimate that behaves worse on this kind of data performs better on data of a different kind.
The overall better results for class N are due to the fact, that there is a bias in the automatically labeled training data towards class N: the percentage of class F stories in the training data is only about 30%, while it is about 40% in the gold standard.

To sum up, the best statistical classifiers are able to classify 76% of the data correctly and precision and recall values, as well as F-measure values for the two classes are ranging from 66% to 85% with the lower values for class F and the higher values for class N. If we combine the automatic labeling with the statistical classifiers then the best combined estimate is able to classify 80% of the data correctly. Considering the average pairwise agreement for manual classification of the gold standard, which is 83.8%, the statistical classifiers perform reasonably well.

Table 6: Evaluation of statistical classifiers and combination of statistical classifiers with automatic labeling against the gold standard.

evaluation measure	results for k nearest neighbor, $k = 16$		results for neural network, 1 hidden neuron		results for *AdaBoost*, 100 stumps	
	SC	**AL-SC**	**SC**	**AL-SC**	**SC**	**AL-SC**
Error	28.75%	20.00%	24.00%	24.25%	25.75%	25.75%
class F, precision	60.34%	71.88%	67.00%	66.83%	66.31%	66.67%
class F, recall	87.20%	84.15%	81.70%	81.10%	75.61%	74.39%
class F, F-measure	71.32%	77.53%	73.62%	73.28%	70.66%	70.32%
class N, precision	87.12%	87.50%	85.00%	84.58%	81.22%	80.65%
class N, recall	60.17%	77.12%	72.03%	72.03%	73.31%	74.15%
class N, F-measure	71.18%	81.98%	77.98%	77.80%	77.06%	77.26%

6. Learning of clues for non-future-directedness

In this section, we describe a method for semi-automatic acquisition of clues for class N and class F from labeled training data. One possible way of acquiring new clues would be to apply information theoretical metrics, such as smoothed pointwise information or information gain. These metrics can be used to measure how indicative a particular word is of a specific category. They have in common that each word is considered separately when its weight according to the metric is determined. We rather opted for a method that determines the weight of a particular word by considering many other words simultaneously. We assume that this way the significance of a particular word for a specific category is easier to recognize. Therefore, we use a particular type of neural network, namely a perceptron, which has only one hidden neuron, to learn new clues. If a perceptron is applied to the automatically created training data, the text features which are discriminative of future-directedness and non-future-directedness can automatically be identified by looking at their associated weights computed during the training phase of the perceptron: text features with a strong positive weight are discriminative of class N, whereas features with a strong negative weight are discriminative of class F. This property of the perceptron can be exploited to learn clues for class N and to extend the list of clues for class F.

In our first experiments on learning new clues we used the Bag-of-Words representation of the 37,703 automatically labeled Reuters documents (see section 5) and allowed the perceptron to train at most 2 hours. The resulting ranking of the 500 most frequent token types was not suited as a basis for a substantial extension of the clue lists, because this ranking contained a high number of proper names. Yet, it provided important insights regarding the role of auxiliary verbs in non-future-directed stories: among the token types with a strong positive weight we detected a number of auxiliary verb forms used to form past and perfect tense in German. In addition, the modal verb forms from our clue set for class F were encountered among the token types with a strong

negative weight. We therefore included the observed auxiliary verb forms in the clue set for class N, see list in appendix D.

Because our primary goal was the acquisition of a wide range of lexical clues for class N, the Bag-of-Words representation was subsequently replaced by a bag-of-features representation. We used three different feature sets, consisting of the 500 most frequent noun lemmas, the 500 most frequent verb lemmas, and the 500 most frequent adjective/adverb lemmas. With each of these bag-of-features representations we trained a perceptron and limited the number of Backpropagation steps made by each perceptron to 30,000. Alternatively, we could have used Rosenblatt's perceptron training algorithm (see, e.g., section 9.1 in [39]), but we did not use this probably more efficient algorithm, since we already had implemented the Backpropagation algorithm. Table 7 shows the 20 top-ranked verbs according to their associated positive weight, which has been computed by a perceptron fed with the 500 most frequent (full) verb lemmas as input data. The third column indicates the results of manually reviewing the verb list.

Table 7: Top 20 verbs indicative of class N as ranked by the perceptron.

weight	verb	verb passes linguistic test
1.1771	*festnehmen* 'arrest'	no
1.1296	*aufhalten* 'stay'	no
1.1287	*entfhren* 'kidnap'	no
1.1269	*verurteilen* 'convict'	no
1.1164	*erobern* 'conquer'	no
1.0785	*angreifen* 'attack'	no
1.0736	*bestreiten* 'deny'	yes
1.0129	*demonstrieren* 'demonstrate against'	no
1.0100	*gestehen* 'admit'	yes
0.9854	*treiben* e.g. 'to bull the market'	no
0.9690	*fragen* 'ask'	yes
0.9328	*verleihen* 'award'	no
0.9288	*landen* 'land'	no
0.9281	*vorwerfen* 'accuse'	yes
0.8906	*blockieren* 'block'	yes
0.8860	*schieen* 'shoot'	no
0.8827	*ermitteln* 'determine, find out'	yes
0.8763	*beschreiben* 'describe, specify'	yes
0.8691	*vergleichen* 'compare'	yes
0.8642	*entdecken* 'discover'	yes

We decided to use a semi-automatic process, mainly because we wanted to get rid of the noisy verbs which necessarily show up when mining of clues is performed fully automatically. The presence of noisy items in the clue list results in a decrease in precision of the automatic labeling procedure. In checking the verb ranking, we were looking for verbs that are similar to the verb clues for class F regarding their linguistic properties, i.e. we were looking for verbs which are able to take clausal complements and which can be used to mark a proposition or action as non-future-directed. Concentrating on such clues makes manual checking easier and more efficient, because the verbs can be checked by applying specific linguistic tests. Otherwise, when also selecting other verbs as

clues, manual checking is much harder, since for many verbs the corresponding sample sentences in the news stories have to be considered in order to decide which verbs to select as clues. As it turned out, there are indeed verb clues for class N which take clausal complements and which mark an action or proposition as non-future-directed. Table 8 shows the broad semantic classes of these verbs together with the corresponding linguistic tests.

Since the perceptron does not distinguish between verbs taking clausal complements and other verbs, the ranking shown in table 7 also contains quite a few verbs from the crime/violence/military actions-domain, for example, *festnehmen* 'arrest', *entführen* 'kidnap', *angreifen* 'attack', *schießen* 'shoot'. It is characteristic of news stories that news about this domain are simply reported, rather than being marked as expected, intended, or even predicted.

Table 8: Linguistic tests for clues for class N.

Verb Class (sample verbs)	Sample Test
REACTION (*bestreiten* 'deny', *gestehen* 'admit') PERCEPTION (*sehen* 'see', *hren* 'hear')	⋆"They verbed an event taking place in the future."
START/END (*blockieren* 'block', *beenden* 'finish')	"They verbed a transaction." ⋆"They verbed their expectations."
ANALYSIS (*entdecken* 'discover', *beschreiben* 'describe') REPORT (*berichten* 'report', *verffentlichen* 'publish')	⋆"They verbed the future."
SUCCEED (*gelingen* 'succeed', *scheitern* 'fail')	"The action verbed."
EVENT (*sich ereignen* 'happen')	"An event verbed."

The whole process of perceptron training and subsequent manual filtering resulted in a set of 98 clues for class N, consisting of 69 verbs, 9 nouns, and 20 adjectives or adverbs, see list in appendix E. Analogously, the clue set for class F has been extended resulting in additional 10 verbs, 8 nouns, and 12 adjectives or adverbs, which are underlined in appendices A and C.

6.1. Improvement of automatic labeling

The labeling algorithm described in section 4.1 had to be slightly modified in order to incorporate the new clues for class N. Now the algorithm counts the number of both future-directed and non-future-directed clues in a given news story, thus computing two relative clue frequencies c_f and c_n corresponding to class F and class N. Then c_f and c_n are compared to four different threshold parameters, $t_{f,1}$, $t_{f,2}$, $t_{n,1}$, and $t_{n,2}$:

if $(c_f > t_{f,1})$ and $(c_n <= t_{n,1})$ then $class := F$

elsif $(c_f <= t_{f,2})$ and $(c_n > t_{n,2})$ then $class := N$

else $class :=< unknown >$

The optimal values for the parameters $t_{f,i}$ and $t_{n,i}$ have been determined automatically by evaluating a wide range of possible parameter combinations on a training set of 200 news stories, the same set which has been used in section 4.1. It turned out, that for class F the value of $t_{n,1}$ is irrelevant: class F can be assigned by checking the number of future-directed clues alone, regardless how high the number of clues for class N is in a given news story. In contrast, for class N to be assigned, the number of clues for class N has to exceed the number of clues for class F. We chose two parameter combinations based on the evaluation results on the training set: one parameter combination with particular high precision of about 95% (**hp**), and another with particular high recall of about 60% and precision not below 80% (**hr**). The different versions of automatically labeled training data created on the basis of these two parameter combinations have been used to answer the question, which version works best in combination with statistical classifiers. Table 9 shows precision and recall of the modified labeling algorithm for **hp** and **hr**, evaluated against the gold standard.

Table 9: Using clues for class F and class N for automatic labeling

evaluation measure	results for parameters **hp** $t_{f,1} = 0.07$, $t_{n,1} = 2.0$ $t_{f,2} = 0.005$, $t_{n,2} = 0.015$	results for parameters **hr** $t_{f,1} = 0.045$, $t_{n,1} = 2.0$ $t_{f,2} = 0.025$, $t_{n,2} = 0.035$
class F, precision	95.45%	79.07%
class F, recall	25.61%	62.20%
class F, F-measure	40.38%	69.63%
class N, precision	94.52%	90.23%
class N, recall	29.24%	66.53%
class N, F-measure	44.66%	76.59%

Comparing the results for **hp** with the results presented in Table 5 in section 4.2, precision for class N has increased by 3%, while recall for class N has decreased only by 2%. Precision for both classes is similar (around 95%), as well as recall, ranging from 26% for class F to 29% for class N.

Considering the results for **hr**, precision for class F is now (evaluated on the gold standard) slightly below 80%, whereas precision for class N is remarkably high at 90%. There is a considerable increase in recall for both classes: class F recall is now 62%, class N recall 67%.

Automatic labeling was performed with the parameters **hp** and **hr**, resulting in two sets of training data consisting of 40,930 and 85,523 news stories which corresponds to 35,22% and 73,59% of all Reuters documents, respectively. The percentage of class F documents in these two sets is similar, namely 41% for **hp** and 43% for **hr**. This is a significant improvement, because the proportion of class F documents in the training data is now as high as in the gold standard.

6.2. Improvement of statistical classification

To begin with, we consider two classifiers which use the prediction of the automatic labeling with parameters **hp** and **hr**, respectively, combined with a guessing step where the class of a document not automatically labeled yet is simply set to class N, which is the majority class in the gold standard. Table 10 shows error, precision, recall and F-measure for these two classifiers.

Table 10: Automatic labeling combined with guessing: evaluation against the gold standard.

evaluation measure	results for automatic **hp**-labeling and guessing	results for automatic **hr**-labeling and guessing
Error	31.00%	22.25%
class F, precision	95.45%	79.07%
class F, recall	25.61%	62.20%
class F, F-measure	40.38%	69.63%
class N, precision	65.73%	77.12%
class N, recall	99.15%	88.56%
class N, F-measure	79.05%	82.45%

As we see from Table 10, the error of the automatic labeling with parameter **hr** combined with guessing is better than the error of the best statistical classifier in section 5 and only slightly worse than the error of the best combined classifier in section 5.

Then we apply the three statistical classifiers to the improved training data. On the **hp**-based training set we did not achieve improvements in terms of error or F-measure, therefore Table 11 shows only the results for the **hr**-based training set.

Table 11: Improved automatic labeling with parameters **hr**: evaluation of statistical classifiers and combination of statistical classifiers with automatic labeling against the gold standard.

evaluation measure	results for k nearest neighbor, $k = 16$		results for neural network, 2 hidden neurons		results for *AdaBoost*, 100 stumps	
	SC	**AL-SC**	**SC**	**AL-SC**	**SC**	**AL-SC**
Error	24.75%	18.75%	20.75%	18.75%	22.00%	19.25%
class F, precision	66.17%	74.32%	71.20%	74.31%	71.35%	73.77%
class F, recall	81.10%	82.93%	82.92%	82.93%	77.44%	82.32%
class F, F-measure	72.88%	78.39%	76.62%	78.39%	74.27%	77.81%
class N, precision	84.42%	87.10%	86.60%	87.10%	83.33%	86.64%
class N, recall	71.19%	80.08%	76.70%	80.08%	78.39%	79.66%
class N, F-measure	77.24%	83.44%	81.35%	83.44%	80.79%	83.00%

Looking at Table 11 we observe that the errors of all classifiers have de-

creased. In particular, the combination of automatic labeling and statistical classifiers now yields significantly better results than the application of statistical classifiers alone, and the best combined classifier is now able to classify 81.75% of the data correctly. As in section 5, the k nearest neighbor classifier as single classifier has the highest error rate, but achieves one of the best results when it is combined with the automatic labels.

7. Discussion

7.1. Related Work

A high-precision labeling method based on presence vs. absence of clues has been described before in subjectivity vs. objectivity classification by [27], [28]. Our approach to classification of future-directedness is most similar to their work. In [27], they describe a bootstrapping process that learns lexical-syntactic subjectivity patterns. It starts with the application of a high-precision classifier for subjective versus objective sentences: the high precision classifier uses lists of lexical subjectivity clues and makes its decision based on the presence or absence of these clues in a given sentence. In the next step, a pattern-extraction algorithm uses the labeled sentences to learn linguistically richer clues, namely lexical-syntactic subjectivity patterns, which are then fed back both to the high-precision classifier and to the pattern-extraction algorithm. Later, [28] report on results of using the automatically labeled sentences as training data for a Naive Bayes classifier.

Besides the fact, that we have been experimenting with document-level classification, rather than sentence-level classification, there are two other significant differences between our work and the work described in [27] and [28]. First, the list of non-future-directed lexical items proposed by the perceptron is manually reviewed in order to acquire linguistically motivated clues, whereas the pattern-extraction algorithm used in [27] works fully automatically. Second, our learning procedure, which is also part of a bootstrapping process, is able to acquire not only further future-directed clues, but also new clues for the complement class, namely non-future-directed documents.

Our particular choice of lexical clues is similar to previous work on subjectivity identification and analysis with respect to the lexico-syntactic richness of the features used: the subjectivity clues used in [27] include manually developed lexical semantic classes taken from [19] among other sources. [43] use a semi-automatically built lexicon of fine-grained semantic features of appraising adjectives and their modifiers to classify movie reviews. The work of [46] on sentiment extraction from online text documents describes the use of sentiment clues in combination with a sentiment pattern database containing lexico-syntactic patterns. The work of [16] describing text mining techniques for collecting domain-dependent evaluative attribute-value-expressions is based on the use of syntactically defined cooccurrence patterns combined with a number of seed attributes and values. [24] report on the extraction of fine-grained features and associated opinions together with their polarity from parsed reviews.

There are, however, two important differences between our clues and previously used clues in subjectivity identification and analysis: First, the verbal clues used here are rather frequently occurring verbs which seem to be largely domain-independent; there are, for example, no especially infrequently occurring terms like the ones which are indicative of subjective expressions, see [27], [44]. Second, our particular choice of verbal clues is highly resource-driven, as lexical-semantic classes of verbs containing future-directed verbs are freely available not only for German, but also for English ([19], [17]). Therefore the transfer of the approach presented in this article to English is presumably straightforward.

As for the automatic acquisition of clues for text classification, there is much previous work on the acquisition of clues or clue expressions in the domain of subjectivity and opinion identification and classification. Two kinds of approaches can be distinguished: either there is prior linguistic knowledge, be it lexical resources, seed terms, or syntactic extraction rules, which is exploited (see [27], [46] [16], [6], [24], [43]), or the acquisition is based on the use of manually labeled training data, see [44], [13]. The acquisition procedure presented in this article is more similar to the latter approach, as it can be applied to labeled training data in general, be it manually labeled or automatically. The use of a perceptron in this context is new to the best of our knowledge. However, the use of a linear model to rank attributes via their weights is a common technique in attribute selection (see, e.g., [45])

7.2. Future Work

In developing lexical clues for future-directedness and non-future-directedness we have concentrated on clues with independent linguistic motivation. Since these clues happen to be largely domain-independent, we presume that they can equally well be used for automatic labeling of texts from a different domain, and we plan to apply the classification tools described here to web content next.

Although the results of our experiments have shown that approximate verb clue frequencies including verb forms in past tense are sufficient to recognize future-directed news stories with high precision, we consider to include a more complete morphological analysis of the text, as well as a shallow syntactic analysis in the future. On the basis of such detailed linguistic preprocessing we could not only accurately determine tense and voice of verbs, but extend our classification method to sentence-level classification, which might be more suitable when dealing with Web content.

Moreover, it would be interesting to transfer the clues proposed for German to English which should be feasible due to the availability of lexical resources in English similar to the ones used here (see [19] and [17]).

Since the results we achieved for pairwise annotator agreement leave room for improvement, due to the document-level annotation task, it is also possible to explore another definition of the gold standard in the future, where a third class NEUTRAL is included which is assigned to news stories which have been annotated as mixed by the annotators.

There are other text classification problems where lexical-semantic classes like the ones presented here can also be used as a discriminant. There are, for

example, lexical-semantic classes of verbs which could be used as clues for opinion vs. non-opinion pieces, objective vs. non-objective pieces, or, speculative vs. non-speculative pieces. Hence the use of a broader range of lexical-semantic classes in the context of various finer-grained text classification problems is another possible issue in future research.

Finally, the perceptron-based learning procedure we have presented in section 6 can be applied to tackle other, related issues in text classification as well. For example, it can be applied

- to learn clues for objective language given a set of automatically labeled training data as in [27];
- to adapt an existing collection of clue words to a new domain;
- to learn clue words for a specific text classification task from scratch, given a set of manually annotated documents.

We would like to thank two anonymous reviewers for many very helpful comments.

References

[1] Artstein, Ron and Massimo Poesio. 2005. Bias decreases in proportion to the number of annotators. In Gerhard Jaeger, Paola Monachesi, Gerald Penn, James Rogers, and Shuly Wintner (eds.), Proceedings of FG-MoL 2005, pages 141-150. Edinburgh, August 2005.

[2] Brandow, Ronald, Karl Mietze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

[3] Devroye, Luc, Lszl Gyrfi, and Gbor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition.* Applications of Mathematics, Springer, New York.

[4] Eckle-Kohler, Judith. 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora.* PhD. thesis, Logos-Verlag, Berlin.

[5] Freund, Y. and R. Schapire. 1997. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences* 55:119–139.

[6] Gamon, Michael and Anthony Aue. 2005. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64, Ann Arbor, Michigan.

[7] Glance, Natalie, Matthew Hurst, and Takashi Tomokiyo. 2004. BlogPulse: Automated Trend Discovery for Weblogs. In *Proceedings of WWW2004*, pages, New York.

[8] Gyrfi, Laszlo, Michael Kohler, Adam Krzyżak, and Harro Walk. 2002. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, Springer, New York.

[9] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. Springer Series in Statistics, Springer, New York.

[10] Hurst, Matthew and Kamal Nigam. 2004. Retrieving Topical Sentiments from Online Document Collections. In *Document Recognition and Retrieval XI, Proceedings of SPIE, Vol. 5296*, pages 27–34, San Jose, California.

[11] Jackendoff, Ray. 1995. The Conceptual Structure of Intending and Volitional Action. In H. Campos and P. Kempchinsky, editors, *Evolution and Revolution in Linguistic Theory: Studies in Honor of Carlos P. Otero*. Georgetown University Press, Washington, pages 198–227.

[12] Kim, Soo-Min and Eduard Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of Conference on Computational Linguistics (COLING-04)*, pages 1367–1373, Geneva, Switzerland.

[13] Kim, Soo-Min and Eduard Hovy. 2005. Automatic Detection of Opinion Bearing Words and Sentences. In Compan*ion Volume to the* Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05), pages 61–66, Jeju Island, Republic of Korea.

[14] Kim, Soo-Min and Eduard Hovy. 2006. Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY.

[15] Klavans, Judith L. and Min-Yen Kan. 1998. Role of Verbs in Document Analysis. In *Proceedings of COLING/ACL*, pages 680–686, Montreal, Canada.

[16] Kobayashi, Nozomi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi (NEC), and Toshikazu Fukushima (NEC). 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 584–589, Sanya City, Hainan Island, China.

[17] Korhonen, Anna and Ted Briscoe. 2004. Extended Lexical-Semantic Classification of English Verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, pages 38–45, Boston, MA.

[18] Korhonen, Anna, Yuval Krymolowski, and Zvika Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics.*, pages 64–71, Sapporo, Japan.

[19] Levin, Beth. 1993. *English Verb Classes and Alternations.* The University of Chicago Press, Chicago.

[20] Lin, Chin-Yew and Eduard Hovy. 1997. Identifying topics by position. In *Proc. of the 5th Applied Natural Language Processing Conference (ANLP-97)*, pages 283–290.

[21] Lin, Wei-Hao, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. In *Proceedings of CoNLL-2006*, pages 109–116, New York, USA.

[22] Merlo, Paola and Suzanne Stevenson. 2001. Automatic Verb Classification based on Statistical Distribution of Argument Structure. *Computational Linguistics*, 27(3):373–408.

[23] Pang, Bo, 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, Philadelphia, PA, USA.

[24] Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. In *Proc. of HLT/EMNLP 2005*, pages 339–346, Vancouver, Canada.

[25] Reuters Corpus. 2000. Volume 1: English Language, 1996-08-20 to 1997-08-19. Release date: 2000-11-03, Format version: 1. http://trec.nist.gov/data/reuters/reuters.html.

[26] Reuters Corpus. 2000. Volume 2: Multilingual Corpus, 1996-08-20 to 1997-08-19. Release date: 2000-05-31, Format version: 1. http://trec.nist.gov/data/reuters/reuters.html.

[27] Riloff, Ellen and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 105–112, Sapporo, Japan.

[28] Riloff, Ellen, Janyce Wiebe, and William Phillips. 2005. Exploiting Subjectivity Classification to Improve Information Extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05).*

[29] Rubin, Victoria L., Noriko Kando, and Elizabeth D. Liddy. 2005. Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text.* Springer, Dordrecht, The Netherlands.

[30] Rudanko, Juhani. 1989. *Complementation and Case Grammar: A Syntactic and Semantic Study of Selected Patterns of Complementation in Present-Day English.* State University of New York Press, Albany, N.Y.

[31] Rudanko, Juhani. 1996. *Prepositions and Complement Clauses: A Syntactic and Semantic Study of Verbs Governing Prepositions and Complement Clauses in Present-Day English.* State University of New York Press, Albany, N.Y.

[32] Saint-Dizier, Patrick. 1999. Verb Semantic Classes in French. In Patrick Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases.* Kluwer Academic Publishers, Dordrecht, pages 139–170.

[33] Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. Guidelines fr das Tagging deutscher Textcorpora mit STTS. Technical report, IMS Stuttgart, SfS Tbingen.

[34] Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

[35] Schulte im Walde, Sabine and Chris Brew. 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.

[36] Schulte im Walde, Sabine. 2004. Induction of Semantic Classes for German Verbs. In Stefan Langer and Daniel Schnorbusch, editors, *Semantik im Lexikon*. Gunter Narr Verlag, Tbingen, pages 61–88.

[37] Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417–424, Philadelphia, PA.

[38] Turney, Peter and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

[39] Vapnik, Vladimir N. 1998. Statistical learning theory. Wiley.

[40] Vzquez, Gloria, Ana Fernndez, Irene Castelln, and María Antonia Martí. 2000. Classificacin Verbal: Alternancias de Ditesis. Number 3 in Quaderns de Sintagma. Universitat de Lleida.

[41] Vendler, Zeno. 1972. *Res Cogitans.* Cornell University Press, 1972.

[42] Vendler, Zeno. 1980. Telling the Facts. In John R. Searle, Ferenc Kiefer, and Manfred Bierwisch, editors, *Speech Act Theory and Pragmatics.* D. Reidel Publishing Company, Dordrecht, Holland, pages 273–290.

[43] Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proc. of ACM Fourteenth Conference on Information and Knowledge Management (CIKM)*, pages 625–631, Bremen, Germany.

[44] Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics* 30(3):277–308.

[45] Witten, Ian H. and Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[46] Yi, Jeonghee, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM '03*, pages 427–434, Washington, DC, USA.

[47] Yu, Hong and Vassileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 129–136, Sapporo, Japan.

[48] Zhu, Xiaojin. 2005. Semi-Supervised Learning Literature Survey. Technical report number 1530, Computer Sciences, University of Wisconsin-Madison.

Appendix A. Lists of German clues taking to-infinitives

Lexemes which are underlined have been added to our given set of clues in section 6, based on the application of a Perceptron.

Predicates marking actions as future-directed

***refuse*-class**
sich weigern, Weigerung, Skrupel, bereit sein, Bereitschaft

***consider*-class**
erwgen, Erwgung, Möglichkeit, denkbar, *schwanken, ausloten, unschlssig*

***intend*-class**
Anreiz, Antrieb, Ehrgeiz, Motivation, Verlockung, Drang, beabsichtigen, Absicht, vorhaben, Vorhaben, Vorsatz, Wille, Entschlossenheit, verzichten, vermeiden *anstreben, Bestreben, sich etw. vornehmen, trachten nach, abzielen auf, anpeilen, vorsehen, hinarbeiten auf, hinwirken auf, eintreten fr, kmpfen um,* durchsetzen *einwilligen, Einwilligung, anhalten zu, Ansto, erforderlich, notwendig, nötig, dringend, brauchen*
Additional related adjectives: anstehend, angestrebt

***try*-class**
versuchen, sich entschlie en, Entschlu

***plan/agree*-class**
abstimmen, sich einigen, Vereinbarung, vereinbart, vorgesehen, vormerken, planen, Plan, geplant, Programm, Konzept, Projekt beraten, diskutieren

***promise/urge*-class**
aufrufen zu, Aufruf, anmahnen, auffordern, Aufforderung, beauftragen, beschwren, drngen, verpflichten erlauben, Erlaubnis, zustimmen, ermöglichen, ermahnen, Ermahnung, ermuntern, ermutigen, Rat, ersuchen, mahnen, Mahnung, nahelegen, anbieten, Angebot, androhen, appellieren an, Appell, drohen, Drohung, warnen vor, Warnung, flehen, fordern, Forderung, beantragen, pldieren fr, propagieren, werben fr, Diskussion versprechen, Versprechen, geloben, warnen vor, Warnung, zusagen, Zusage, zusichern, Zusicherung

***suggest*-class**
bitten, Bitte, empfehlen, Empfehlung, vorschlagen, Vorschlag, anflehen

***decide*-class**
ankndigen, Ankndigung, beschlie en, Beschlu , entscheiden, Entscheidung
Additional related verbs and adjectives: *verabschieden*, *beschlossen*

Predicates marking propositions as future-directed

***cause*-class**
fhren zu, bewirken, hinauslaufen auf, beitragen, Voraussetzung

***hope*-class**
ausgehen von, Annahme, erwarten, mutma en, Mutmaung, befrchten, Befrchtung, frchten, Furcht, sich sorgen, Sorge, erhoffen, hoffen, Hoffnung, wnschen, Wunsch, setzen auf, vertrauen auf
Additional related adjectives: *erwartet*

***wait*-class**
rechnen mit, voraussehen, einkalkulieren, warten, zittern, wetten, Wette

***predict*-class**
bangen, schtzen, Schtzung, Einschtzung, abschtzen, spekulieren, Spekulation, prognostizieren, Prognose, prophezeien, Prophezeiung, vorhersagen, Vorhersage, voraussagen, Voraussage

Appendix B. Modal verb forms indicating future-directedness

soll, solle, sollen, sollte, sollten, mu , msse, mssen, mte, mten, will, wolle, wollen, drfte, drften, knnte, knnten

Appendix C. Temporal adverbs and temporal expressions indicating future-directedness

- *weiterhin, voraussichtlich, knftig, zuknftig, bald, sptestens, frhestens, kommend*
- *in Zukunft, ab Anfang, ab Mitte, ab Ende, ab dem, bis am*

Appendix D. Auxiliary verb forms indicating non-future-directedness

hat, hatte, wurde, wurden, worden, war, waren, ist, sind, gewesen

Appendix E. Lists of German clues for non-future-directedness

REACTION
reagieren, Reaktion, verstehen, hinnehmen, bewerten, werten, beurteilen, feiern, bekennen, gestehen, sich erinnern, bestreiten, dementieren, widersprechen, beschuldigen, vorwerfen, Vorwurf, unerwartet, auffllig, berraschend

PERCEPTION
hren, sehen, beobachten

REPORT
berichten, melden, mitteilen, angeben, Aussage, vorlegen, verffentlichen, bekanntgeben, besttigen, aussprechen, lauten, verlauten

ANALYSIS
entdecken, Hinweis, ermitteln, Ermittlung, erkennen, verbuchen, registrieren, belaufen, verzeichnen, berechnen, beziffern, ergeben, Durchschnitt, erzielen, betragen, ausfallen, notieren, beschreiben, bezeichnen, feststellen, stimmen, tatschlich, korrekt, falsch, anders, scheinen, fragen, fehlen, entsprechen, bedeuten, beziehen, zurckfhren, Zusammenhang, verursachen, Grund, vergleichen, Vergleich, vergleichbar, unterschiedlich

START/END
auslsen, einleiten, beenden, enden, stoppen, abbrechen, blockieren, unterbrechen, aussetzen

SUCCEED
gelingen, scheitern

EVENT
sich ereignen

Temporal adverbs and adjectives *damals, bisherig, vergangen, zuvor, ehemalig, vorher, anfnglich, mittlerweile, inzwischen, immer, nie*