Automatic Focus Annotation: Bringing Formal Pragmatics Alive in Analyzing the Information Structure of Authentic Data

Ramon Ziai **Detmar Meurers**
Collaborative Research Center 833
University of Tübingen
{rziai,dm}@sfs.uni-tuebingen.de

Abstract

Analyzing language in context, both from a theoretical and from a computational perspective, is receiving increased interest. Complementing the research in linguistics on discourse and information structure, in computational linguistics identifying discourse concepts was also shown to improve the performance of certain applications, for example, Short Answer Assessment systems (Ziai and Meurers, 2014).

Building on the research that established detailed annotation guidelines for manual annotation of information structural concepts for written (Dipper et al., 2007; Ziai and Meurers, 2014) and spoken language data (Calhoun et al., 2010), this paper presents the first approach automating the analysis of focus in authentic written data. Our classification approach combines a range of lexical, syntactic, and semantic features to achieve an accuracy of 78.1% for identifying focus.

1 Introduction

The interpretation of language is well known to depend on context. Both in theoretical and computational linguistics, discourse and information structure of sentences are thus receiving increased interest: attention has shifted from the analysis of isolated sentences to the question how sentences are structured in discourse and how information is packaged in sentences analyzed in context.

As a consequence, a rich landscape of approaches to discourse and information structure has been developed (Kruijff-Korbayová and Steedman, 2003). Among these perspectives, the Focus-Background dichotomy provides a particularly valuable structuring of the information in a sentence in relation to the discourse. (1) is an example question-answer pair from Krifka and Musan (2012, p. 4) where the focus in the answer is marked by brackets.

(1) Q: *What did John show Mary?*
A: *John showed Mary* $[\![$*the PICtures*$]\!]_F$.

In the answer in (1), the NP *the pictures* is focussed and hence indicates that there are alternative things that John could show Mary. It is commonly assumed that focus here typically indicates the presence of alternative *denotations* (denotation focus, Krifka and Musan 2012, p.8), making it a semantic notion. Depending on the language, different devices are used to mark focus, such as prosodic focus marking or different syntactic constructions (e.g. clefts). In this paper, we adopt a notion of focus based on alternatives, as advanced by Rooth (1992) and more recently, Krifka and Musan (2012), who define focus as indicating “the presence of alternatives that are relevant for the interpretation of linguistic expressions” (Krifka and Musan, 2012, p. 7). Formal semantics has tied the notion of alternatives to an explicit relationship between questions and answers called Question-Answer Congruence (Stechow, 1991), where the idea is that an answer is congruent to a question if both evoke the same set of alternatives. Questions can thus be seen as a way of making alternatives explicit in the discourse, an idea also taken up by the Question-Under-Discussion (QUD) approach (Roberts, 2012) to discourse organization.

Complementing the theoretical linguistic approaches, in the last decade corpus-based approaches started exploring which information structural notions can reliably be annotated in what kind of language data. While the information status (Given-New) dimension can be annotated successfully (Riester et al., 2010; Nissim et al., 2004) and even automated (Hempelmann et al., 2005; Nissim, 2006; Cahill and Riester, 2012), the inter-annotator agreement results for Focus-Background (Ritz et al., 2008; Calhoun et al., 2010) show that it is difficult to obtain high levels of agreement, especially due to disagreement about the extent or size of the focused unit.

More recently, Ziai and Meurers (2014) showed that for data collected in task contexts including explicit questions, such as answers to reading comprehension questions, reliable focus annotation is possible. In addition, an option for externally validating focus annotation was established by showing that such focus annotation improves the performance of Short Answer Assessment (SAA) systems. Focus enables the system to zoom in on the part of the answer addressing the question instead of considering all parts of the answer as equal.

In this paper, we want to build on this strand of research and develop an approach for automatically identifying focus in authentic data including explicit question contexts. In contrast to Calhoun (2007) and Sridhar et al. (2008), who make use of prosodic properties to tackle the identification of focus for content words in spoken language data, we target the analysis of written texts.

We start in section 2 by discussing relevant related work before introducing the gold standard focus annotation we are using as foundation of our work in section 3. Section 4 then presents the different types of features used for predicting which tokens form a part of the focus. In section 5 we employ a supervised machine learning setup to evaluate the perspective and specific features in terms of the ability to predict the gold standard focus labeling. Building on these intermediate results and the analysis thereof in section 6, in section 7 we then present two additional feature groups which lead to our final focus detection model. Finally, section 8 explores options for extrinsically showing the value of the automatic focus annotation for the automatic meaning assessment of short answers. It confirms that focus analysis pays off when aiming to generalize assessment to previously unseen data and contexts.

2 Previous Approaches

There is only a very small number of approaches dealing with automatically labeling information structural concepts.[1] Most approaches related to detecting focus automatically almost exclusively center on detecting the 'kontrast' notion in the English Switchboard corpus (Calhoun et al., 2010). We therefore focus on the Switchboard-based approaches here.

The availability of the annotated Switchboard corpus (Calhoun et al., 2005, 2010) sparked interest in information-structural categories and enabled several researchers to publish studies on detecting focus. This is especially true for the Speech Processing community, and indeed many approaches described below are intended to improve computational speech applications in some way, by detecting prominence through a combination of various linguistic factors. Moreover, with the exception of Badino and Clark (2008), all approaches use prosodic or acoustic features.

All approaches listed below tackle the task of detecting 'kontrast' (as focus is called in the Switchboard annotation) automatically on various subsets of the corpus using different features and classification approaches. For each approach, we therefore report the features and classifier used, the data set size as reported by the authors, the (often very high) majority baseline for a binary distinction between 'kontrast' and background, and the best accuracy obtained. If available in the original description of the approach, we also report the accuracy obtained without acoustic and prosodic features.

Calhoun (2007) investigated how focus can be predicted through what she calls "prominence structure". The essential claim is that a "focus is more likely if a word is more prominent than expected given its syntactic, semantic and discourse properties". The classification experiment is based on 9,289 words with a 60% majority baseline for the 'background' class. Calhoun (2007) reports 77.7% for a combination of prosodic, syntactic and semantic features in a logistic regression model. Without the prosodic and acoustic features, the accuracy obtained is at 74.8%. There is no information on a separation between training and test set, likely due to the setup of the study being geared towards determining relevant factors in predicting focus, not building a focus prediction model for a real application case. Relatedly, the approach uses only gold-standard annotation already available in the corpus as the basis for features, not automatic annotation.

Sridhar et al. (2008) use lexical, acoustic and part-of-speech features in trying to detect pitch accent, givenness and focus. Concerning focus, the work attempts to extend Calhoun (2007)'s analysis to "understand what prosodic and acoustic dif-

[1] For a broader perspective of computational approaches in connection with information structure, see Stede (2012).

ferences exist between the focus classes and background items in conversational speech". 14,555 words of the Switchboard corpus are used in total, but filtered for evaluation later to balance the skewed distribution between 'kontrast' and 'background'. With the thus obtained random baseline of 50%, Sridhar et al. (2008) obtain 73% accuracy when using all features, which again drops only slightly to 72.95% when using only parts of speech. They use a decision tree classifier to combine the features in 10-fold cross-validation for training and testing.

Badino and Clark (2008) aim to model contrast both for its role in analyzing discourse and information structure, and for its potential in speech applications. They use a combination of lexical, syntactic and semantic features in an SVM classifier. No acoustic or prosodic features are employed in the model. In selecting the training and testing data, they filter out many 'kontrast' instances, such as those triggered across sentence boundaries, those above the word level, and those not sharing the same broad part of speech with the trigger word. The resulting data set has 8,602 instances, of which 96.8% are 'background'. Badino and Clark (2008) experiment with different kernel settings for the SVM and obtain the best result of 97.19% using a second-order polynomial kernel, and leave-one-out testing.

In contrast to all approaches above, we target the analysis of written texts, for which prosodic and acoustic information is not available, so we must rely on lexis, syntax and semantics exclusively. Also, the vast majority of the approaches discussed make direct use of the manually annotated information in the corpus they use in order to derive their features. While this is a viable approach when the aim is to determine the relevant factors for focus detection, it does not represent a real-life case where annotated data often unavailable. In our focus detection model, we only use automatically determined annotation as the basis for our features for predicting focus.

Since our approach also makes use of question properties, it is also worth mentioning that there are a number of approaches on Answer Typing as a step in Question Answering (QA) approaches in order to constrain the search space of possible candidate answers and improve accuracy. While earlier approaches such as Li and Roth (2002) used a fixed set of answer types for classifying factoid questions, other approaches such as Pinchak and Lin (2006) avoid assigning pre-determined classes to questions and instead favor a more data-driven label set. In more recent work, Lally et al. (2012) use a sophisticated combination of deep parsing, lexical clues and broader question labels to analyze questions.

3 Data

The present work is based on the German CREG corpus (Ott et al., 2012). CREG contains responses by American learners of German to comprehension questions on reading texts. Each response is rated by two teaching assistants with regard to whether it answers the question or not. While many responses contain ungrammatical language, the explicit questions in CREG generally make it possible to interpret responses. More importantly for our work, they can be seen as Questions Under Discussion and thus form an ideal foundation for focus annotation in authentic data.

As a reference point for the automatic detection of focus, we used the CREG-ExpertFocus data set (De Kuthy et al., 2016) containing 3,187 student answers and 990 target answers (26,980 words in total). It was created using the incremental annotation scheme described in Ziai and Meurers (2014), where annotators first look at the surface question form, then determine the set of alternatives, and finally mark instances of the alternative set in answers. De Kuthy et al. (2016) report substantial agreement in CREG-ExpertFocus ($\kappa \geq .7$) and provide an adjudicated gold standard, which thus presents a high-quality basis for training our focus detection classifier.

4 Focus Detection Model

As described in section 3 above, focus was marked in a span-based way in the data set used: each instance of focus starts at a specific word and ends at another word. Since in principle any part of speech can be focused, we cannot constrain ourselves to a pre-defined set of markables for automatic classification. We therefore conceptualized the task of automatic focus detection on a per-word level: for each word in an answer, as identified by the OpenNLP tokenizer and sentence segmenter[2], the classifier needs to decide whether it is an instance of *focus* or *background*. Besides the choice of

[2] `http://opennlp.apache.org`

classification algorithm, the crucial question naturally is the choice of linguistic features, which we turn to next.

4.1 Features

Various types of linguistic information on different linguistic levels can in principle be relevant for focus identification, from morphology to semantics. We start by exploring five groups of features, which are outlined below. In section 7, we discuss two more groups designed to address specific problems observed with the initial model.

Syntactic answer properties (SynAns) A word's part-of-speech and syntactic function are relevant general indicators with respect to focus: since we are dealing with meaning alternatives, the meaning of e.g. a noun is more likely to denote an alternative than a grammatical function word such as a complementizer or article.

Similarly, a word in an argument dependency relation is potentially a stronger indicator for a focused alternative in a sentence than a word in an adjunct relation. We therefore included two features: the word's **part-of-speech** tag in the STTS tag set (Schiller et al., 1995) determined using TreeTagger (Schmid, 1994), and the **dependency relation to the word's head** in the Hamburg dependency scheme (Foth et al., 2014, p. 2327) determined using MaltParser (Nivre et al., 2007) as features in our model.

Question properties The question constitutes the direct context for the answer and dictates its information structure and information requirements to fulfill. In particular, the type of *wh*-phrase (if present) of a question is a useful indicator of the type of required information: a *who*-question, such as 'Who rang the doorbell?', will typically be answered with a noun phrase, such as 'the milkman'. We identified **surface question forms** such as *who*, *what*, *how* etc. using a regular expression approach developed by Rudzewitz (2015) and included them as features. Related to question forms, we also extracted the question word's **dependency relation to its head**, analogous to the answer feature described above.

Surface givenness As a rough and robust approximation to information status, we add a boolean feature indicating the **presence of the current word in the question**. We use the lemmatized form of the word as determined by TreeTagger (Schmid, 1994).

Positional properties Where a word occurs in the answer or the question can be relevant for its information structural status. It has been observed since Halliday (1967) that given material tends to occur earlier in sentences (here: answers), while new or focused content tends to occur later. We encode this observation in three different features: the **position of the word in the answer** (normalized by sentence length), the **distance from the finite verb** (in words), and the **position of the word in the question** (if it is given).

Conjunction features To explicitly tie answer properties to question properties, we explored different combinations of the features described above. Specifically, we encoded the **current word's POS depending on the question form**, and the **current word's POS depending on the *wh*-word's POS**. To constrain the feature space and get rid of unnecessary distinctions, we converted the answer word's POS to a coarse-grained version before computing these features, which collapses all variants of determiners, pronouns, adjectives/adverbs, prepositions, nouns and verbs into one label, respectively.[3]

5 Intrinsic Evaluation

5.1 Setup

To employ the features described above in an actual classifier, we trained a logistic regression model using the WEKA toolkit (Hall et al., 2009). We also experimented with other classification algorithms such as SVMs, but found that they did not offer superior performance for this task. The data set used consists of all expert focus annotation available (3,187 student answers, see section 3), with the exception of the answers occurring in the extrinsic evaluation test set we use in section 8, which leaves a total of 2,240 student answers with corresponding target answers and questions. We used 10-fold cross-validation on this data set to experiment and select the optimal model for focus detection.

[3] For a list (in German) of the full tag set, see `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html`

5.2 Results

Table 1 lists the accuracies[4] obtained for our different feature groups, as well as three baselines: a POS baseline, following Sridhar et al. (2008), a baseline that only includes the simple givenness feature, and the majority baseline. The majority class is *focus*, occurring in 58.1% of the 26,980 cases (individual words).

	Accuracy for		
Feature set	*focus*	*backgr.*	both
Majority baseline	100%	0%	58.1%
Givenness baseline	81.5%	42.5%	65.1%
POS baseline	89.2%	39.6%	68.4%
SynAns	82.8%	50.3%	69.2%
+ Question	83.8%	53.1%	70.9%
+ Given	84.8%	62.0%	74.8%
+ Position	84.9%	66.5%	77.2%
+ Conjunction	85.2%	66.7%	**77.4%**

Table 1: Initial focus detection model

We can see that each feature group incrementally adds to the final model's performance, with particularly noticeable boosts coming from the givenness and positional features. Another clear observation is that the classifier is much better at detecting*focus* than *background*, possibly also due to the skewedness of the data set. Note that performance on *background* increases also with the addition of the 'Question' feature set, indicating the close relation between the set of alternatives introduced by the question and the focus selecting from that set, even though our approximation to computationally determining alternatives in questions is basic. It is also clear that the information intrinsic in the answers, as encoded in the 'SynAns' and 'Position' feature sets, already provides significant performance benefits, suggesting that a classifier trained only on these features could be trained and applied to settings where no explicit questions are available.

6 Qualitative Analysis

In order to help explain the gap between automatic and manual focus annotation, let us take a step back from quantitative evaluation and examine a few characteristic examples in more detail.

Figure 1 shows a case where a *why*-question is answered with an embedded 'weil' (because) clause. The classifier successfully marked 'weil' and the end of the clause as *focus*, but left out the pronoun 'es' (it) in the middle, presumably because pronouns are given and often not focused in other answers. We did experiment with using a sequence classification approach in order to remedy such problems, but it performed worse overall than the logistic regression model we presented in section 4. We therefore suggest that in such cases, a global constraint stating that *why*-questions are typically answered with a full clause would be a more promising approach, combining knowledge learned bottom-up from data with top-down linguistic insight.

In Figure 2, we can see two different problems. One is again a faulty gap, namely the omission of the conjunction 'und' (and). The other is the focus marking of the word 'AG' (corporation) in the beginning of the sentence: since the question asks for an enumeration of the institutions that form a corporation, marking 'AG' as focused is erroneous. This problem likely occurs often with nouns because the classifier has learned that content words are often focused. Moreover, the surface givenness feature does not encode that 'AG' is in fact an abbreviation of 'Aktiengesellschaft' and therefore given. It would thus be beneficial to extend our analysis of givenness beyond surface identity, a direction we explore in the next section.

Finally, Figure 3 presents a case where an enumeration is marked correctly, including the conjunctive punctuation in between, showing that cases of longer foci are indeed within reach for a word-by-word focus classifier.

7 Extending the Model

Based on our analysis of problematic cases outlined in the previous section, we explored two different avenues for improving our focus detection model, which we describe below.

7.1 Distributional Givenness

We have seen in section 5.2 that surface-based givenness is helpful in predicting focus. However, it clearly has limitations, as for example synonymy cannot be captured on the surface. We also exemplified one such limitation in Figure 2. In order to overcome these limitations, we implemented an approach based on distributional semantics. This avenue is motivated by the fact that Ziai et al. (2016) have shown Givenness modeled

[4] We show per-class and overall accuracies, the former is also known as recall or true positive rate.

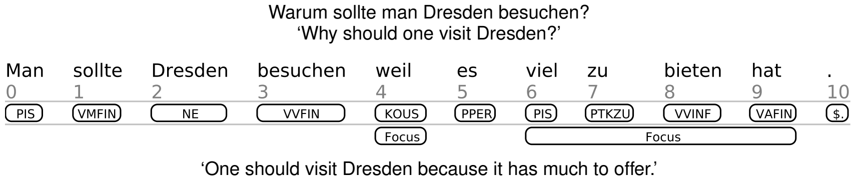

Warum sollte man Dresden besuchen?
'Why should one visit Dresden?'
Man sollte Dresden besuchen weil es viel zu bieten hat .
0 1 2 3 4 5 6 7 8 9 10
PIS VMFIN NE VVFIN KOUS PPER PIS PTKZU VVINF VAFIN $.
Focus
Focus
'One should visit Dresden because it has much to offer.'

Figure 1: Focus with a faulty gap in between

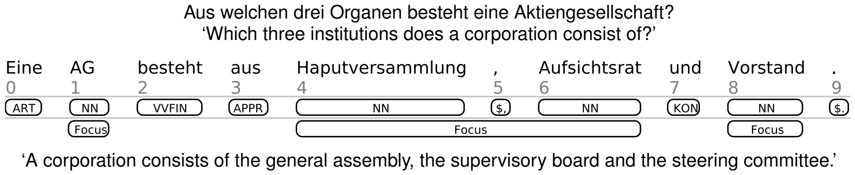

Aus welchen drei Organen besteht eine Aktiengesellschaft?
'Which three institutions does a corporation consist of?'
Eine AG besteht aus Haputversammlung , Aufsichtsrat und Vorstand .
0 1 2 3 4 5 6 7 8 9
ART NN VVFIN APPR NN $, NN KON NN $.
Focus
Focus
Focus
'A corporation consists of the general assembly, the supervisory board and the steering committee.'

Figure 2: Focus with a faulty outlier (and a faulty gap)

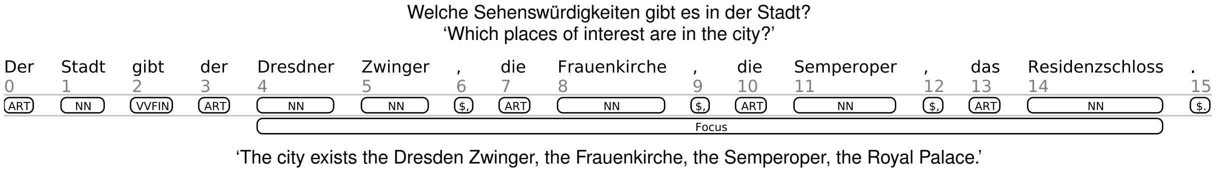

Welche Sehenswürdigkeiten gibt es in der Stadt?
'Which places of interest are in the city?'
Der Stadt gibt der Dresdner Zwinger , die Frauenkirche , die Semperoper , das Residenzschloss .
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
ART NN VVFIN ART NN NN $, ART NN $, ART NN $, ART NN $.
Focus
'The city exists the Dresden Zwinger, the Frauenkirche, the Semperoper, the Royal Palace.'

Figure 3: Enumeration with correct focus

as distributional similarity to be helpful for SAA at least in some cases. We used the word vector model they derived from the DeWAC corpus (Baroni et al., 2009) using word2vec's continuous bag-of-words training algorithm with hierarchical softmax (Mikolov et al., 2013). The model has a vocabulary of 1,825,306 words and uses 400 dimensions for each.

Having equipped ourselves with a word vector model, the question arises how to use it in focus detection in such a way that it complements the positive impact that surface-based givenness already demonstrates. Rather than using an empirically determined (and hence data-dependent) empirical threshold for determining givenness as done by Ziai et al. (2016), we here use raw cosine similarities[5] as features and let the classifier assign appropriate weights to them during training. Concretely, we calculate **maximum, minimum and average cosine between the answer word and the question words**. As a fourth feature, we calculate the **cosine between the answer word and the additive question word vector**, which is the sum of the individual question word vectors.

7.2 Constituency-based Features

Another source of evidence we wanted to exploit is constituency-based syntactic annotation. So far, we have worked with part-of-speech tags and dependency relations as far as syntactic representation is concerned. However, while discontinuous focus is possible, focus as operationalized in the scheme by Ziai and Meurers (2014) most often marks an adjacent group of words, a tendency that our word-based classifier did not always follow, as exemplified by the cases in Figures 1 and 2. Such groups very often correspond to a syntactic phrase, so constituent membership is likely indicative in predicting the focus status of an individual word. Similarly, the topological field (Höhle, 1986) identifying the major section of a sentence in relation to the clausal main verb is potentially relevant for a word's focus status.

Cheung and Penn (2009) present a parsing model that demonstrates good performance in determining both topological fields and phrase structure for German. The model is trained on the TüBa-D/Z treebank (Telljohann et al., 2004), whose rich syntactic model encodes topological fields as nodes in the syntax tree itself. Following Cheung and Penn (2009), we trained an updated version of their model using the current version of the Berkeley Parser (Petrov and Klein, 2007) and release 10 of the TüBa-D/Z.[6]

Based on the new parsing model, we integrated two new features into our focus detection model:

[5] We normalize cosine similarity as cosine distance to obtain positive values between 0 and 2: $dist = 1 - sim$

[6] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

the **direct parent constituent node of a word** and **the nearest topological field node of a word**.

7.3 Final Results

Table 2 shows the impact of the new feature groups discussed above.

	Accuracy for		
Feature set	*focus*	*backgr.*	both
Majority baseline	100%	0%	58.1%
Givenness baseline	81.5%	42.5%	65.1%
POS baseline	89.2%	39.6%	68.4%
Initial model (sec. 5.2)	85.2%	66.7%	77.4%
+ dist. Givenness	84.7%	68.0%	77.7%
+ constituency	84.8%	68.7%	**78.1%**

Table 2: Final focus detection performance

While the improvements may seem modest quantitatively, they show that the added features are well-motivated and do make an impact. Overall, it is especially apparent that the key to better performance is reducing the number of false positives in this data set: while the accuracy for focus stays roughly the same, the one for background improves steadily with each feature set addition.

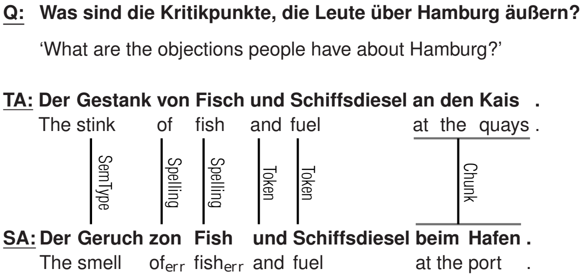
8 Extrinsic Evaluation

Complementing the intrinsic evaluation above, in this section we demonstrate how focus can be successfully used to improve performance in an authentic CL task, namely Short Answer Assessment (SAA).

8.1 Setup

It has been pointed out that evaluating the annotation of a theoretical linguistic notion only intrinsically is problematic because there is no non-theoretical grounding involved (Riezler, 2014). Therefore, besides a comparison to the gold standard, we also evaluated the resulting annotation in a larger computational task, the automatic meaning assessment of short answers to reading comprehension questions. Here the goal is to decide, given a question (Q) and a correct target answer (TA), whether the student answer (SA) actually answers the question or not. An example from Meurers et al. (2011) is shown in Figure 4.

We used the freely available CoMiC system (Comparing Meaning in Context, Meurers et al. 2011) as a testbed for our experiment. CoMiC is an alignment-based system operating in three stages:


Q: Was sind die Kritikpunkte, die Leute über Hamburg äußern?
'What are the objections people have about Hamburg?'

TA: Der Gestank von Fisch und Schiffsdiesel an den Kais .
The stink of fish and fuel at the quays .

SA: Der Geruch zon Fish und Schiffsdiesel beim Hafen .
The smell of$_{err}$ fish$_{err}$ and fuel at the port .

Figure 4: Short Answer Assessment example

1. Annotating linguistic units (words, chunks and dependencies) in student and target answer on various levels of abstraction
2. Finding alignments of linguistic units between student and target answer based on annotation (see Figure 4)
3. Classifying the student answer based on number and type of alignments (see Table 3), using a supervised machine learning setup

Feature	Description
1. Keyword Overlap	Percent of dependency heads aligned (relative to target)
2./3. Token Overlap	Percent of aligned target/student tokens
4./5. Chunk Overlap	Percent of aligned target/student chunks (as identified by OpenNLP[3])
6./7. Triple Overlap	Percent of aligned target/student dependency triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments resolved using PMI-IR (Turney, 2001)
10. Type Match	Percent of token alignments resolved using GermaNet hierarchy (Hamp and Feldweg, 1997)
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments sharing same GermaNet synset
13. Variety of Match (0-5)	Number of kinds of token-level alignments (features 8–12)

Table 3: Standard features in the CoMiC system

In stage 2, CoMiC integrates a simplistic approach to givenness, excluding all words from alignment that are mentioned in the question. We transferred the underlying method to the notion of focus and implemented a component that excludes all non-focused words from alignment, resulting

[3] http://opennlp.apache.org/

in alignments between focused parts of answers only. The hypothesis is that the alignment of focused elements in answers adds information about the quality of the answer with respect to the question, leading to a higher answer classification accuracy.

We experimented with two different settings involving the standard CoMiC system and a focus-augmented variant: i) using standard CoMiC with the givenness filter by itself as a baseline, and ii) augmenting standard CoMiC by additionally producing a focus version of each classification feature in Table 3. In each case, we used WEKA's *k*-nearest-neighbor implementation for CoMiC, following positive results by Rudzewitz (2016).

We use two test sets randomly selected from the CREG-5K data set (Ziai et al., 2016), one based on an 'unseen answers' and one based on an 'unseen questions' test scenario, based on the methodology of (Dzikovska et al., 2013): in 'unseen answers', the test set can contain answers to the same questions already part of the training set (but not the answers themselves), whereas in 'unseen questions' both questions and answers are new in the test set. In order to arrive at a fair and generalizable testing setup, we removed all answers from the CREG-5K training set that also occur in the CREG-ExpertFocus set used to train our focus detection classifier. This ensures that neither the focus classifier nor CoMiC have seen any of the test set answers before.

The resulting smaller training set contains 1606 student answers, while the test sets contain 1002 (unseen answers) and 1121 (unseen questions), respectively.

8.2 Results

Table 4 summarizes the results for the different CoMiC variants and test sets in terms of accuracy in classifying answers as *correct* vs. *incorrect*. 'Standard CoMiC' refers to the standard CoMiC system and '+Focus' refers to the augmented system using both feature versions. For reference on what is possible with Focus information, we provide the results of the oracle experiment by De Kuthy et al. (2016), even though the test setup and data setup are slightly different. In addition to our two test sets introduced above, we tested the systems on the training set using 10-fold cross validation. We also provide the majority baseline of the respective data set along with the majority class.

One can see that in general, the focus classifier seems to introduce too much noise to positively impact classification results. The standard CoMiC system outperforms the focus-augmented version for the cross validation case and the 'unseen answers' set. This is in contrast to the experiments reported by De Kuthy et al. (2016) using manual focus information, where the augmented system clearly outperforms all other variants. This shows that while focus information is clearly useful in Short Answer Assessment, it needs to be reliable enough to be of actual benefit. Recall also that the way we use focus information in CoMiC implies a strong commitment: only focused words are aligned and included in feature extraction, which does not produce the desired result if the focus information is not accurate. A possible way of remedying this situation would be to use focus as an extra feature or less strict modifier of existing features. There is thus room for improvement both in the automatic detection of focus and its use in extrinsic tasks.

However, one result stands out encouragingly: in the 'unseen questions' case, the focus-augmented version beats standard CoMiC, if only by a relatively small margin. This shows that even automatically determined information structural properties provide benefits when more concrete information, in the form of previously seen answers to the same questions, is not available. Our classifier thus successfully transfers general knowledge about focus to new question material.

9 Conclusion

We presented the first automatic focus detection approach for written data, and the first such approach for German. The approach uses a rich feature set including abstractions to grammatical notions (parts of speech, dependencies), word order aspects captured by a topological field model of German, an approximation of Givenness and the relation between material in the answer and that of the question word.

Using a word-by-word classification approach that takes into account both syntactic and semantic properties of answer and question words, we achieve an accuracy of 78.1% on a data set of 26,980 words in 10-fold cross validation. The focus detection pipeline developed for the experiment is freely available to other researchers.

Complementing the intrinsic evaluation, we

Test set	Instances	Majority baseline	CoMiC	+Focus
Oracle experiment reported by De Kuthy et al. (2016) on CREG-ExpertFocus				
leave-one-out	3187	51.0% *(correct)*	83.2%	85.6%
10-fold CV	1606	54.4% *(correct)*	**83.2%**	82.3%
Unseen answers	1002	51.3% *(correct)*	**80.6%**	80.5%
Unseen questions	1121	51.1% *(incorrect)*	77.4%	**78.4%**

Table 4: CoMiC results on different test sets using standard and focus-augmented features

provide an extrinsic evaluation of the approach as part of a larger CL task, the automatic content assessment of answers to reading comprehension questions. We show that while automatic focus detection does not yet improve content assessment for answers similar to the ones previously seen, it does provide a benefit in test cases where the questions and answers are completely new, i.e., where the system needs to generalize beyond the specific cases and contexts previously seen.

Contextualizing our work, one can see two different strands of research in the automatic analysis of focus. In comparison to Calhoun (2007) and follow-up approaches, who mainly concentrate on linking prosodic prominence to focus in dialogues, we do not limit our analysis to content words, but analyze every word of an utterance. This is made feasible due to the explicit task context we have in the form of answers to reading comprehension questions. We believe this nicely illustrates two avenues for obtaining relevant evidence on information structure: On the one hand, there is evidence obtained bottom-up through the data such as the rich information on prominence in spoken language data such as the corpus used by Calhoun (2007). On the other hand, there is top-down evidence from the task context, which sets up expectations about what is to be addressed for the current question under discussion. Following the QUD research strand, the approach presented in this paper could be scaled up beyond explicit question-answer pairs: De Kuthy et al. (2018) spell out an explicit analysis of text in terms of QUDs and show that it is possible to annotate explicit QUDs with high inter-annotator agreement. Combined with an automated approach to question generation, it could thus be possible to recover implicit QUDs from text and subsequently apply our current approach to any text, based on an independently established, general formal pragmatic analysis.

Finally, the qualitative analysis we exemplified is promising in terms of obtaining valuable insights to be addressed in future work. For example, the analysis identified faulty gaps in focus marking. In future work, integrating insights from theoretical linguistic approaches to focus and the notion of focus projection established there (cf., e.g., De Kuthy and Meurers 2012) could provide more guidance for ensuring contiguity of focus domains.

Acknowledgements

We would like to thank Kordula De Kuthy and the anonymous reviewers for detailed and helpful comments on different versions of this paper. This work has been funded by the Deutsche Forschungsgemeinschaft through Collaborative Research Center 833.

References

Leonardo Badino and Robert A. J. Clark. 2008. Automatic labeling of contrastive word pairs from spontaneous spoken English. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*. pages 101–104. https://doi.org/10.1109/SLT.2008.4777850.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation* 3(43):209–226. http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in german. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 232–236. http://aclweb.org/anthology/W12-1632.

Sasha Calhoun. 2007. Predicting focus through prominence structure. In *Proceedings of Interspeech*. Antwerp, Belgium. http:

//www.cstr.inf.ed.ac.uk/downloads/publications/2007/calhounIS07.pdf.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation* 44:387–419. http://link.springer.com/article/10.1007\%2Fs10579-010-9120-1.

Sasha Calhoun, Malvina Nissim, Mark Steedman, and Jason Brenier. 2005. A framework for annotating information structure in discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 45–52. http://aclweb.org/anthology/W/W05/W05-0307.

Jackie Chi Kit Cheung and Gerald Penn. 2009. Topological field parsing of german. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Association for Computational Linguistics, Morristown, NJ, USA, pages 64–72. http://aclweb.org/anthology/P09-1008.

Kordula De Kuthy and Detmar Meurers. 2012. Focus projection between theory and evidence. In Sam Featherston and Britta Stolterfoht, editors, *Empirical Approaches to Linguistic Theory – Studies in Meaning and Structure*, De Gruyter, volume 111 of *Studies in Generative Grammar*, pages 207–240. http://purl.org/dm/papers/dekuthy-meurers-11.html.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. Qud-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, JP.

Kordula De Kuthy, Ramon Ziai, and Detmar Meurers. 2016. Focus annotation of task-based data: a comparison of expert and crowd-sourced annotation in a reading comprehension corpus. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC)*. Portorož, Slovenia, pages 3928–3934. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1083_Paper.pdf.

Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, volume 7 of *Interdisciplinary Studies on Information Structure*. Universitätsverlag Potsdam, Potsdam, Germany. http://www.sfb632.uni-potsdam.de/publications/isis07.pdf.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 263–274. http://aclweb.org/anthology/S13-2045.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The hamburg dependency treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 2326–2333. http://www.lrec-conf.org/proceedings/lrec2014/pdf/860_Paper.pdf.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*. volume 11, pages 10–18.

Michael Halliday. 1967. Notes on Transitivity and Theme in English. Part 1 and 2. *Journal of Linguistics* 3:37–81, 199–244.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. http://aclweb.org/anthology/W97-0802.

Christian F. Hempelmann, David Dufty, Philip M. McCarthy, Arthur C. Graesser, Zhiqiang Cai, and Danielle S. McNamara. 2005. Using LSA to automatically identify Givenness and Newness of noun phrases in written discourse. In B. G. Bara, L. Barsalou, and M. Bucciarelli, editors, *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. Erlbaum, Stresa, Italy, pages 941–949. https://doi.org/10.1.1.116.5716.

Tilman N. Höhle. 1986. Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne, editor, *Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985*, Niemeyer, Tübingen, pages 329–340. Bd. 3.

Manfred Krifka and Renate Musan. 2012. Information structure: overview and linguistic issues. In Manfred Krifka and Renate Musan, editors, *The Expression of Information Structure*, De Gruyter Mouton, Berlin/Boston, volume 5 of *The Expression of Cognitive Categories*, pages 1–43.

Ivana Kruijff-Korbayová and Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information (Introduction to the Special Issue)* 12(3):249–259.

Adam Lally, John M. Prager, Michael C. McCord, Branimir K. Boguraev, Siddharth Patwardhan, James Fan, Paul Fodor, and Jennifer Chu-Carroll. 2012. Question analysis: How Watson reads a clue. *IBM Journal of Research and Development* 56(3/4):2:1–14.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan, pages 1–7. http://aclweb.org/anthology/C02-1150.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*. Edinburgh, pages 1–9. http://aclweb.org/anthology/W11-2401.pdf.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Emprical Methods in Natural Language Processing*. Sydney, Australia.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th Conference on Language Resources and Evaluation*. Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/638.pdf.

Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(1):1–41. http://w3.msi.vxu.se/~nivre/papers/nle07.pdf.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Benjamins, Amsterdam, Hamburg Studies in Multilingualism (HSM), pages 47–69. https://benjamins.com/\#catalog/books/hsm.14.05ott.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pages 404–411.

Christopher Pinchak and Dekang Lin. 2006. A probabilistic answer type model. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Lingustics (EACL)*. pages 393–400.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/764_Paper.pdf.

Stefan Riezler. 2014. On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics* 40(1):235–245.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, pages 2137–2142. http://www.lrec-conf.org/proceedings/lrec2008/pdf/543_paper.pdf.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6):1–69. https://doi.org/10.3765/sp.5.6.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1):75–116.

Björn Rudzewitz. 2015. *Alignment Weighting for Short Answer Assessment*. Bachelor's thesis, University of Tübingen. www.sfs.uni-tuebingen.de/~brzdwtz/resources/BA_Thesis.pdf.

Björn Rudzewitz. 2016. Exploring the intersection of short answer assessment, authorship attribution, and plagiarism detection. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA, pages 235–241. https://aclweb.org/anthology/W16-0527.pdf.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. The Stuttgart-Tübingen Tagset (STTS). Technical report, Universität Stuttgart, Universität Tübingen, Germany. http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Vivek Kumar Rangarajan Sridhar, Ani Nenkova, Shrikanth Narayanan, and Dan Jurafsky. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*. Campinas, Brazil, pages 380–388.

Arnim von Stechow. 1991. Focusing and backgrounding operators. In W. Abraham, editor, *Discourse Particles*, John Benjamins Publishing Co., Amsterdam/Philadelphia, pages 37–84.

Manfred Stede. 2012. Computation and modeling of information structure. In Manfred Krifka and Renate Musan, editors, *The Expression of Information Structure*, De Gruyter Mouton, Berlin/Boston, volume 5 of *The Expression of Cognitive Categories*, pages 363–408.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lissabon.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*. Freiburg, Germany, pages 491–502.

Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2016. Approximating Givenness in Content Assessment through Distributional Semantics. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*. ACL, Berlin, Germany, pages 209–218. http://aclweb.org/anthology/S16-2026.pdf.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*. COLING, ACL, Dublin, Ireland, pages 159–168. http://aclweb.org/anthology/W14-4922.pdf.