Approximating Givenness in Content Assessment through Distributional Semantics

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

Abstract

Givenness (Schwarzschild, 1999) is one of the central notions in the formal pragmatic literature discussing the organization of discourse. In this paper, we explore where distributional semantics can help address the gap between the linguistic insights into the formal pragmatic notion of Givenness and its implementation in computational linguistics.

As experimental testbed, we focus on short answer assessment, in which the goal is to assess whether a student response correctly answers the provided reading comprehension question or not. Current approaches only implement a very basic, surface-based perspective on Givenness: A word of the answer that appears as such in the question counts as GIVEN.

We show that an approach approximating Givenness using distributional semantics to check whether a word in a sentence is similar enough to a word in the context to count as GIVEN is more successful quantitatively and supports interesting qualitative insights into the data and the limitations of a basic distributional semantic approach identifying Givenness at the lexical level.

1 Introduction

Givenness is one of the central notions in the formal pragmatic literature discussing the organization of discourse. The distinction between *given* and *new* material in an utterance dates back at least to Halliday (1967) where *given* is defined as "anaphorically recoverable" and the notion is used to predict patterns of prosodic prominence. Schwarzschild (1999) proposes to define Givenness in terms of the entailment of the existential f-closure between previously mentioned material and the GIVEN expression, hereby also capturing the occurrence of synonyms and hyponyms as given.

On the theoretical linguistic side, a foundational question is whether an approach to Information Structure should be grounded in terms of a Given-New or a Focus-Background dichotomy, or whether the two are best seen as complementing each other. Computational linguistic research on short answer assessment points in the direction of both perspectives providing performance gains (Ziai and Meurers, 2014). On the empirical side, the characteristic problem of obtaining high inter-annotator agreement in focus annotation (Ritz et al., 2008; Calhoun et al., 2010) can be overcome through an incremental annotation process making reference to questions as part of an explicit task context (Ziai and Meurers, 2014; De Kuthy et al., 2016).

In short answer assessment approaches determining whether a student response correctly answers a provided reading comprehension question, the practical value of excluding material that is mentioned in the question from evaluating the content of the answer has been clearly established (Meurers et al., 2011; Mohler et al., 2011). Yet these computational linguistic approaches only implement a very basic, completely surface-based perspective on Givenness: A word of the answer that appears as such in the question counts as GIVEN.

Such a surface-based approach to Givenness fails to capture that the semantic notion of Givenness

i) may be transported by *semantically similar* words,

ii) *entailment* rather than identity is at stake, and

iii) so-called *bridging* cases seem to involve *semantically related* rather than *semantically similar* words.

Computational linguistic approaches to classifying Givenness (Hempelmann et al., 2005; Nissim, 2006; Rahman and Ng, 2011; Cahill and Riester, 2012) have concentrated on the information status of noun phrases, without taking into account other syntactic elements. Furthermore, they do not explicitly make use of similarity and relatedness between lexical units as we propose in this paper. Our approach thus explores a new avenue in computationally determining Givenness.

Theoretical linguistic proposals spelling out Givenness are based on formal semantic formalisms and notions such as logical entailment, type shifting, and existential f-closure, which do not readily lend themselves to extending the computational linguistic approaches. As already alluded to by the choice of words "semantically similar" and "semantically related" above, in this paper we want to explore whether distributional semantics can help address the gap between the linguistic insights into Givenness and the computational linguistic realizations. In place of surface-based Givenness checks, as a first step in this direction we developed an approach integrating distributional semantics to check whether a word in a sentence is similar enough to a word in the context to count as GIVEN.

In section 2, we provide the background on Schwarzschild's notion of Givenness and conceptually explore what a distributional semantic perspective may offer. Section 3 then introduces the application domain of content assessment as our experimental sandbox and the CoMiC system (Meurers et al., 2011) we extended. The distributional model for German used in extending the baseline system is built in section 4. In section 5 we then turn to the experiments we conducted using the system extended with the distributional Givenness component and provide quantitative results. Section 6 then presents the qualitative perspective, discussing examples to probe into the connection between the theoretical linguistic notion of Givenness and its distributional semantic approximation, and where it fails. Finally, section 7 concludes with a summary of the approach and its contribution.

2 Linking Givenness and the distributional semantic perspective

Before turning to the computational realization and a quantitative and qualitative evaluation of the idea, let us consider which classes of data are handled by the theoretical linguistic approach to Givenness and where an approximation of Givenness using distributional semantics can contribute.

Let us first define Givenness according to Schwarzschild (1999, p. 151): an utterance U counts as GIVEN iff it has a salient antecedent A and either i) A and U co-refer or ii) A entails the Existential F-Closure of U. In turn, the Existential F-Closure of U is defined as "the result of replacing F-marked phrases in U with variables and existentially closing the result, modulo existential type shifting" (Schwarzschild, 1999, p. 150).

Schwarzschild uses Givenness to predict where in an utterance the prosodic prominence falls. Consider the question-answer pair in (1), example (12) of Schwarzschild (1999).

(1) John drove Mary's red convertible. What did he drive before that?
A: He drove her BLUE convertible.

Here the prominence does not fall on *convertible* as the rightmost expression answering the question, as generally is the case in English, but instead on the adjective *blue* because the *convertible* is GIVEN and thus is de-accented according to Schwarzschild. With respect to our goal of automatically identifying Givenness, such cases involving **identical lexical material** that is repeated (here: *convertible*) are trivial for a surface-based or distributional semantic approach.

A more interesting case of Givenness involves **semantically similar** words such as synonyms and hypernyms, as exemplified by *violin* and *string instrument* in (2), mentioned as example (7) by Büring (2007).

(2) (I'd like to learn the violin,) because I LIKE string instruments.

The existence of a violin entails the existence of a string instrument, so *string instrument* is GIVEN and deaccented under Schwarzschild's approach. Such examples are beyond a simple surface-based approach to the identification of Givenness and motivate the perspective pursued in this paper: investigating whether a distributional semantic approach to semantic similarity can be used to capture them.

Before tackling these core cases, let us complete the empirical overview of the landscape of cases that the Givenness notion is expected to handle. A relevant phenomenon in this context is **bridging**. It can be exemplified using (3), which is example (29) of Schwarzschild (1999).

(3) a. John got the job.
b. I KNOW. They WANTed a New Yorker.

The part of the formal definitions that is intended to capture the deaccenting of *New Yorker* in a context where *John* is known to be from that city simply refers to salience (Schwarzschild, 1999: "An utterance U counts as GIVEN iff it has a salient antecedent A ..."), which Schwarzschild readily admits is not actually modeled: "Exactly which propositions count as in the background for these purposes remains to be worked out". While beyond the scope of our experiments, approaches computing semantic similarity in more local contexts, such as Dinu and Lapata (2010), may be able to provide an avenue for handling such narrowly contextualized notions of common ground in the evolving, dynamic discourse.

A more straightforward case arises when such bridging examples involve semantic relatedness between expressions that are richly represented in corpora. For example, the fact that Giuliani was the mayor of New York and thus can be identified as semantically related to New Yorker in (4) is within reach of a distributional semantic approach.

(4) a. Giuliani got the job.
b. I KNOW. They WANTed a New Yorker.

When exactly such bridging based on semantically related material results in GIVEN material and its deaccenting, as far as we are aware, has not been systematically researched and would be relevant to explore in the future.

An interesting case related to bridging that adds a further challenge for any Givenness approach is exemplified by (5), originating as example (4) in Büring (2007). The challenge arises from the fact that it does not seem to involve an apparent semantic relation such as entailment – yet the accent falling on *strangle* can only be explained if *butcher* is GIVEN, i.e., entailed by the context.

(5) a. Did you see Dr. Cremer to get your root canal?
b. (Don't remind me.) I'd like to STRANgle the butcher.

The linguistic approaches to Givenness do not formally tackle this since the lexical semantic specification and contextual disambiguation of *butcher* as a particular (undesirable type of) *dentist* is beyond their scope. The fact that *butcher* counts as GIVEN is not readily captured by a general distributional semantic approach either since it is dependent on the specific context and the top-down selection of the meaning of *butcher* as referring to people who brutally go about their job. Distributional semantic approaches distinguishing specific word senses (Iacobacci et al., 2015) could be applicable for extending the core approach worked out in this paper to cover such cases.

Overall, at the conceptual level, a realization of Givenness in terms of distributional semantics can be seen as nicely complementing the theoretical linguistic approach in terms of the division of labor of formal and distributional factors.

3 Content Assessment: Baseline System and Gold Standard Data

To be able to test the idea we conceptually motivated above, we chose short answer assessment as our experimental testbed. The content assessment of reading comprehension exercises is an authentic task including a rich, language-based context. This makes it an interesting real-life challenge for research into the applicability of formal pragmatic concepts such as Givenness. Provided a text and a question, the content assessment task is to determine whether a particular response actually answers the question or not.

In such a setting, the question typically introduces some linguistic material about which additional information is required. The material introduced is usually not the information required in a felicitous answer. For example, in a question such as 'Where was Mozart born?', we are looking for a location. Consequently, in an answer such as 'Mozart was born in Salzburg', we can disregard the words 'Mozart', 'was' and 'born' on account of their previous mention, leaving only the relevant information 'in Salzburg'.

Short answer assessment is thus a natural testbed since the practical value of excluding material that is mentioned in the question from evaluating the content of the answer has been clearly established (Meurers et al., 2011; Mohler et al., 2011) – yet these approaches only integrated a basic surface-based perspective on Givenness. The CoMiC system (Meurers et al., 2011) is freely available, so we used it as baseline approach and proceeded to replaced its surface-based Givenness filter with our distributional semantic approach to Givenness.

3.1 Baseline system

CoMiC is an alignment-based Content Assessment system which assesses student answers by analyzing the quantity and quality of alignment links it finds between the student and the target answer. For content assessment, it extracts several numeric features based on the number and kind of alignments found between non-GIVEN answer parts. The only change we made to the baseline setup is to replace the TiMBL (Daelemans et al., 2007) implementation of k-nearest-neighbors with the WEKA package (Hall et al., 2009), setting k to 5 following the positive results of Rudzewitz (2016).

The CoMiC system we use as baseline for our research employs a surface-based Givenness filter, only aligning tokens not found in the question. The surface-based Givenness filter thus ensures that parts of the answer already occurring in the question are not counted (or could be fed into separate features so that the machine learner making the final assessment can take their discourse status into account).

3.2 Gold-standard content assessment corpus

The data we used for training and testing our extension of the CoMiC system are taken from the CREG corpus (Ott et al., 2012), a task-based corpus consisting of answers to reading comprehension questions written by American learners of German at the university level. It was collected at Kansas University (KU) and The Ohio State University (OSU). The overall corpus includes 164 reading texts, 1,517 reading comprehension questions, 2,057 target answers provided by the teachers, and 36,335 learner answers.

The CREG-5K subset used for the present experiments is an extended version of CREG-1032 (Meurers et al., 2011), selected using the same criteria after the overall, four year corpus collection effort was completed. The criteria include balancedness (same number of correct and incorrect answers), a minimum answer length of four tokens, and a language course level at the intermediate level or above.

4 Creating a distributional model

To model Givenness as distributional similarity, we need an appropriate word vector model. As there is no such model readily available for German, we trained one ourselves.

As empirical basis, we used the DeWAC corpus (Baroni et al., 2009) since it is a large corpus that is freely available and it is already lemmatized, both of which have been argued to be desirable for word vector models. Further preprocessing consisted of excluding numbers and other undesired words such as foreign language material and words the POS tagger had labelled as non-words. The whole corpus was converted to lowercase to get rid of unwanted distinctions between multiple possible capitalizations.

To select an implementation for our purpose, we compared two of the major word vector toolkits currently available, word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). While word2vec is a prediction-based approach that optimizes the probability of a word occurring in a certain context, GloVe is a counting approach based on co-occurrences of words.

We compared the two on the lexical substitution task designed for GermEval 2015 (Miller et al., 2015). The task can be seen as related to recognizing Givenness: deciding what a good substitute for a word in context is requires similar mechanisms to deciding whether the meaning of a word is already present in previous utterances. For GloVe, we used the models trained by Dima (2015), which were also trained on a large German web corpus and were shown to perform well. However, results on the lexical substitution task put both of word2vec's training approaches, continuous bag-of-words (CBOW) and skip-gram, ahead of GloVe using the models previously mentioned, so we continued with word2vec.

Finally, to select the optimal training algorithm for word2vec for our purpose, we again used the GermEval task as a benchmark. We explored both CBOW and skip-gram with negative sampling and hierarchical softmax, yielding four combinations. Among these, CBOW with hierarchical softmax significantly outperformed all other combinations, so we chose it as our training algorithm.

The German model we obtained has a vocabulary of 1,825,306 words and uses 400 dimensions for each, the latter being inspired by Iacobacci et al. (2015).

5 Experiment and Quantitative Results

Now that we have a baseline content assessment system (section 3) and a distributional model for German (section 4) in place, we have all the components to quantitatively and qualitatively evaluate the idea to model Givenness through semantic similarity measures. To do so, we simply replaced the surface-based Givenness filter of the baseline CoMiC system with a distributional-semantics based Givenness filter based on the model described in the previous section. For this we must make concrete, how exactly distributional-semantic distances are used to determine the words in an answer counting as GIVEN.

The parameters to be estimated relate to two different ways one can determine semantic relatedness using word vectors for two words w_1 and w_2:

I. Calculate cosine similarity of w_1 and w_2 and require it to be at least equal to a threshold t.

II. Calculate n nearest words to w_1 and check whether w_2 is among them.

For the first method, one needs to estimate the threshold t, while for the second method one needs to determine how many neighbors to calculate (n). We explored both methods. For the threshold parameter t, we experimented with values from 0.1 to 0.9 in increments of 0.1. For the number of nearest neighbors n, we used a space from 2 to 20 with increments of 2.

To cleanly separate our test data from the data used for training and parameter estimation, we randomly sampled approximately 20% of the CREG-5K data set and set it aside as the final test set. The remaining 80% was used as training set. All parameter estimation was done before running the final system on the test set and using only the training data.

Table 1 shows the results in terms of classification accuracy for 10-fold cross-validation on the training data. The table includes the performance of the system without a Givenness filter as well as with the basic surface-based approach. Training and testing was done separately for the two sub-corpora of CREG-5K corresponding to the universities where they were collected, KU and OSU.

	KU	OSU
# answers	1466	2670
Without Givenness	75.4%	76.7%
Surface Givenness	82.4%	83.0%
Best threshold t	0.3	0.5
Accuracy using t	82.7%	**83.6%**
Best n nearest-words	20	10
Accuracy using n	**83.2%**	**83.6%**

Table 1: Content Assessment results on training set

First, the results confirm that an alignment-based content assessment system such as CoMiC greatly benefits from a Givenness filter, as demonstrated by the big gap in performance between the no-Givenness and surface-Givenness conditions. Second, both the threshold method and the nearest-words method outperform the surface baseline, if only by a small margin.

Turning to the actual testing, we wanted to find out whether the improvements found for the distributional-semantic Givenness filters carry over to the untouched test set. We trained the classifier on the full training set and used the best parameters from the training set. The results thus obtained are summarized in Table 2.

	KU	OSU
# answers	348	654
No Givenness	74.7%	74.2%
Surface Givenness	80.7%	81.2%
Accuracy using t	81.0%	**81.8%**
Accuracy using n	**81.9%**	81.0%

Table 2: Content Assessment results on test set

We can see that results on the test set are generally lower, but the general picture for the test set is the same as what we found for the 10-fold CV on the training data: Surface-based-Givenness easily outperforms the system not employing a Givenness filter, and at least one of the systems employing a distributional semantic Givenness filter (marginally) outperforms the surface-based method.

Interestingly, the two data sets seem to differ in terms of which relatedness method works best for recognizing Givenness: while the threshold method works better for OSU, the n-nearest-words method is the optimal choice for the KU data set. This may be due to the fact that the OSU data set is generally more diverse in terms of lexical variation and thus presents more opportunities for false positives, i.e., words that are somewhat related but should not be counted as given. Such cases are better filtered out using a global threshold. The KU data set, on the other hand, contains less variation and hence profits from the more local n-nearest-words method, which always returns a list of candidates for any known word in the vocabulary, no matter whether the candidates are globally very similar or not.

6 Qualitative Discussion

While the quantitative results provide a useful ballpark measure of how well a Givenness filter based on distributional semantics performs and that it can improve the content assessment of reading comprehension questions, the relatively small and heterogeneous nature of the data set for a complex task such as the content assessment of reading comprehension means that such quantitative results by themselves are best interpreted cautiously. For the conceptual side of our proposal, it is more interesting to see whether semantic similarity can adequately capture the different types of Givenness that we discussed in section 2.

6.1 Successfully identifying Givenness through distributional semantics

To illustrate how exactly the Givenness filter in the CoMiC system ensures that only the material that is not already present in the question is aligned for assessing the similarity of a student and a target answer, let us start by taking a look at a simple example from CREG where the answers repeat lexical material from the question, as shown Figure 1.

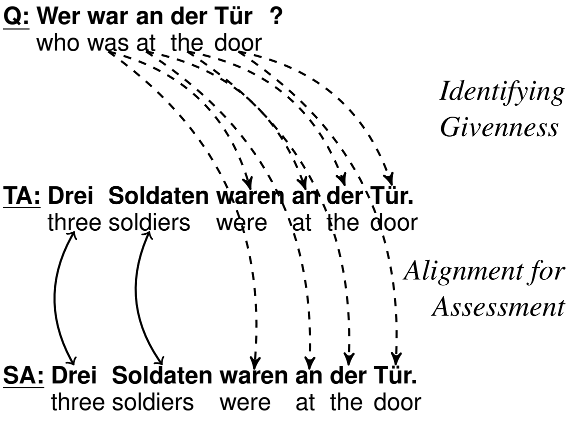
Q: Wer war an der Tür ?
who was at the door
Identifying Givenness
TA: Drei Soldaten waren an der Tür.
three soldiers were at the door
Alignment for Assessment
SA: Drei Soldaten waren an der Tür.
three soldiers were at the door

Figure 1: Simple Givenness alignment

The dotted arrows show which words in the question trigger Givenness marking of which items in the target and the student answer. The solid arrows illustrate the actual alignments between words in the target and the student answer used in the content assessment.

The Givenness filter ensures that the words *waren* (*was*), *an* (*at*), *der* (*the*), and *Tür* (*door*) of the student (SA) and the target (TA) answers are marked as GIVEN with respect to the question and are thus not aligned in order to calculate the similarity of the two answers.

A type of Givenness that a surface-based Givenness filter cannot handle, but that is captured by our distributional similarity approach, occurs in examples where parts of the question are picked up by semantically similar words in the target and student answer. This is illustrated by Figure 2.

The verbs *glaubte* (*believed*) and *meinte* (*thought*) are semantically close enough to the verb *verstand* (*understood*) in the question for them to be identified as GIVEN. They consequently can be excluded from the content assessment of the student answer (SA) in relation to the target answer (TA).

The core idea to use semantic similarity as identified by distributional semantics to identify the words which are GIVEN in a context thus nicely captures real cases in authentic data.

6.2 Overidentifying Givenness

At the same time, there are two aspects of distributional semantics that can also lead to overidentification of Givenness.

Entailment is not symmetric, but semantic similarity and relatedness are The first difficulty arises from the fact that semantic similarity and semantic relatedness are symmetric, whereas the entailment relation used to define Givenness is not. As a result, our distributional semantic model wrongly identifies a word as GIVEN that is more specific than, i.e., a hyponym of the word in the context as illustrated in Figure 3.

The entire NP *praktische Erfahrung im Controlling eines Finanzservice-Unternehmens* (*practical experience in controlling of a financial service company*) consists of new material in both the target answer and the student answer and should thus be aligned for the content assessment of the student answer. But since *Finanzservice-Unternehmen* (*financial service company*) is semantically similar to the noun *Firma* (*company*) occurring in the question, it is marked as GIVEN under the current setting of our distributional similarity approach and incorrectly excluded from the content assessment.

Under the notion of Givenness as defined by Schwarzschild, *Finanzservice-Unternehmen* (*financial service company*) would not count as GIVEN, since the mentioning of *company* in the prior discourse does not entail the existence of a *financial service company*.

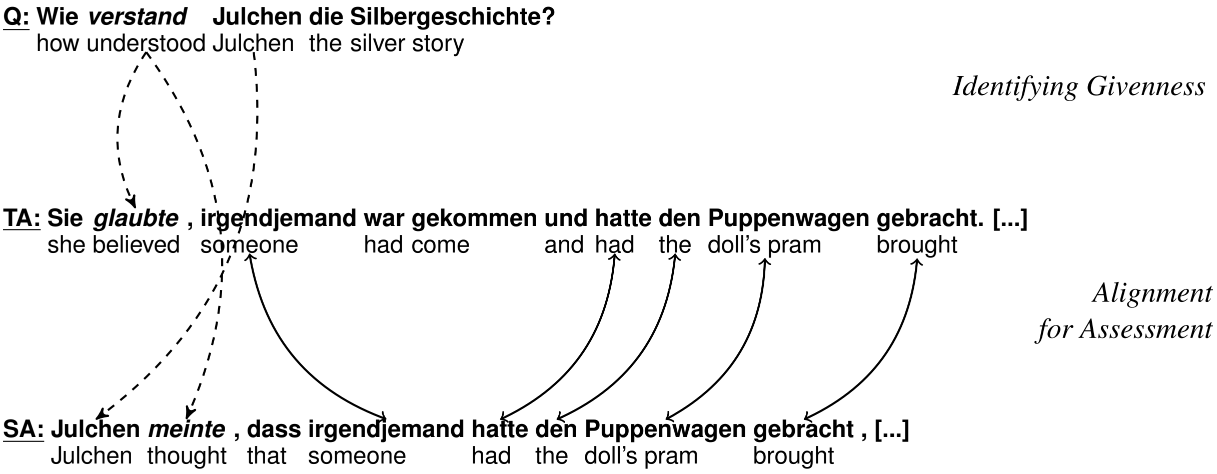
Q: Wie *verstand* Julchen die Silbergeschichte?
how understood Julchen the silver story
Identifying Givenness
TA: Sie *glaubte* , irgendjemand war gekommen und hatte den Puppenwagen gebracht. [...]
she believed someone had come and had the doll's pram brought
Alignment for Assessment
SA: Julchen *meinte* , dass irgendjemand hatte den Puppenwagen gebracht , [...]
Julchen thought that someone had the doll's pram brought

Figure 2: CREG example illustrating Semantic similarity

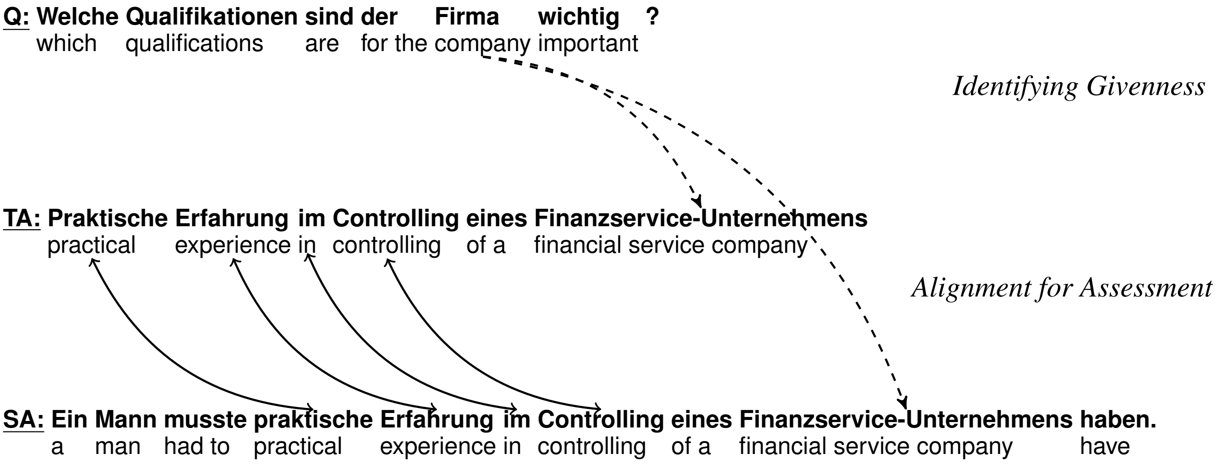
Q: Welche Qualifikationen sind der Firma wichtig ?
which qualifications are for the company important
Identifying Givenness
TA: Praktische Erfahrung im Controlling eines Finanzservice-Unternehmens
practical experience in controlling of a financial service company
Alignment for Assessment
SA: Ein Mann musste praktische Erfahrung im Controlling eines Finanzservice-Unternehmens haben.
a man had to practical experience in controlling of a financial service company have

Figure 3: CREG example illustrating entailment in wrong direction

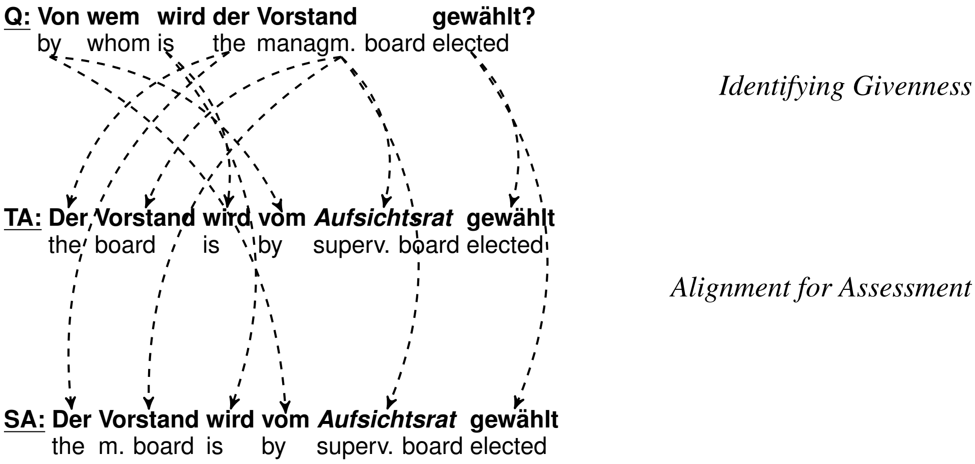
Q: Von wem wird der Vorstand gewählt?
by whom is the managm. board elected
Identifying Givenness
TA: Der Vorstand wird vom *Aufsichtsrat* gewählt
the board is by superv. board elected
Alignment for Assessment
SA: Der Vorstand wird vom *Aufsichtsrat* gewählt
the m. board is by superv. board elected

Figure 4: CREG example illustrating Semantic Relatedness

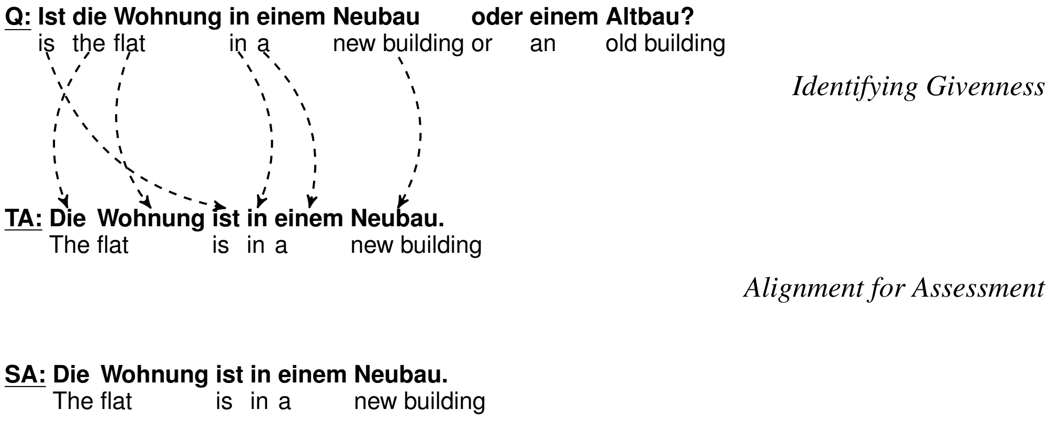
Q: Ist die Wohnung in einem Neubau oder einem Altbau?
is the flat in a new building or an old building
Identifying Givenness
TA: Die Wohnung ist in einem Neubau.
The flat is in a new building
Alignment for Assessment
SA: Die Wohnung ist in einem Neubau.
The flat is in a new building

Figure 5: CREG example illustrating overidentification by Givenness filter

Semantic relatedness is not semantic similarity Second, it is difficult for distributional semantic approaches to distinguish semantic similarity from semantic relatedness (cf., e.g., Kolb, 2009). In the discussion of bridging in section 2 we saw that cases such as (4) could arguably benefit from the use of semantic relatedness to identify Givenness. Yet, allowing all semantic related material to count as GIVEN clearly overestimates what counts as GIVEN and can therefore be deaccented. As a result, our approach wrongly identifies some semantic relatedness cases as Givenness. Consider the semantically related words *Vorstand* (*management board*) and *Aufsichtsrat* (*supervisory board*) in the example shown in Figure 4.

The Givenness filter ensures that the lexical material *der* (*the*), *Vorstand* (*management board*), *wird* (*is*), *gewählt* (*elected*) that is repeated in the answers is marked as GIVEN and thus excluded from the content assessment. But under the current setting of our distributional similarity approach, the noun *Aufsichtsrat* (*supervisory board*) that is semantically related to the noun *Vorstand* (*advisory board*) is also marked as GIVEN and thus excluded from the content assessment. As a consequence all material in the answers is excluded from the alignment and the CoMiC system fails to classify the student answer as a correct answer. A general solution to this kind of misidentification seems to be beyond the scope of an analysis based on the word level – an issue which also turns out to be a problem in another, systematic set of cases, which we turn to next.

Comparing lexical units not enough The Givenness filter under both approaches, surface-based Givenness as well as distributional similarity, sometimes also overidentifies Givenness because the analysis is based on lexical units rather than entailment between sentence meanings. Recall that the way this filter works is to exclude tokens from alignment which are GIVEN in the question. But what if the lexical material required by the question is actually explicitly spelled out as an option by the question itself? This actually happens systematically for alternative questions, where one has to pick one out of an explicitly given set of alternatives. Consider the example in Figure 5, where target and student answer happen to be identical (and for visual transparency only the arcs between question and target answer are shown, not also the identical arcs that link the question and the student answer).

The question asks whether the apartment is in a new or in an old building. Both alternatives are GIVEN in the question, however only one is correct, namely that the apartment is in a new building. The student correctly picked that alternative, but the Givenness filter excludes all material from alignment for content assessment. Hence, classification fails to mark this as a correct answer. As a simple fix, one could integrate an automatic identification of question types and switch off the Givenness filter for alternative questions. More interesting would be an approach that explores when material provided by the question constitutes alternatives in the sense of focus alternatives (Krifka, 2007), from which a selection in the answer should be counted as informative. This essentially would replace the Givenness filter with an approach zooming in to the material in Focus in the answer in the context of the question. At the same time, realizing this idea would require development of an approach automatically identifying Focus, an alternative avenue to pursue in future research.

7 Conclusion

The paper investigated how the formal pragmatic notion of Givenness can be approximated using current computational linguistic methods, and whether this can capture a number of distinct conceptual subcases. We tested the idea in a real-life computational linguistic task with an established external evaluation criterion, content assessment of learner answers to reading comprehension questions.

In place of a surface-based Givenness filter as employed in previous content assessment work, we developed an approach based on distributional semantics to check whether a word in an answer is similar enough to a word in the question to count as GIVEN. The quantitative evaluation confirms the importance of a Givenness filter for content assessment and improved content assessment accuracy for the distributional approach. We experimented with absolute cosine similarity thresholds and with calculating the nearest n words for a candidate word and found that which of the two works better potentially depends on data set characteristics such as lexical diversity.

In the qualitative evaluation, we confirmed that the approximation of Givenness through semantic similarity does indeed capture a number of conceptual cases that a pure surface-based Givenness approach cannot handle, such as bridging-cases involving semantically related words – though this can also lead to over-identification. In future research, integrating more context-sensitive notions of semantic similarity, such as proposed by Dinu and Lapata (2010), may provide a handle on a more narrowly contextualized notion of Givenness in the common ground of discourse participants.

Acknowledgments

We would like to thank Mohamed Balabel for his work on training and selecting the distributional model. Furthermore, we are grateful to the anonymous reviewers of JerSem and *Sem for their comments, which were very helpful in revising the paper.

References

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.

Daniel Büring. 2007. Intonation, semantics and information structure. In Gillian Ramchand and Charles Reiss, editors, *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

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*, pages 232–236. Association for Computational Linguistics.

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.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2007. *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands, July 11. Version 6.0.

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).*

Corina Dima. 2015. Reverse-engineering language: A study on the semantic compositionality of german compounds. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1637–1642, Lisbon, Portugal, September. Association for Computational Linguistics.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, MA, October. Association for Computational Linguistics.

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.

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*, pages 941–949, Stresa, Italy. Erlbaum.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Peter Kolb. 2009. Experiments on the difference between semantic similarity and relatedness. In Kristiina Jokinen and Eckhard Bick, editors, *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 81–88.

Manfred Krifka. 2007. Basic notions of information structure. In Caroline Fery, Gisbert Fanselow, and Manfred Krifka, editors, *The notions of information structure*, volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*, pages 13–55. Universitätsverlag Potsdam, Potsdam.

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*, pages 1–9, Edinburgh, July. ACL.

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.

Tristan Miller, Darina Benikova, and Sallam Abualhaija. 2015. GermEval 2015: LexSub – A shared task for German-language lexical substitution. In *Proceedings of GermEval 2015: LexSub*, pages 1–9, sep.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

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*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1069–1080, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

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*, pages 2137–2142, Marrakech, Morocco.

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.

Roger Schwarzschild. 1999. GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, pages 159–168, Dublin, Ireland. COLING, Association for Computational Linguistics.