Unsupervised Methods for Domain Specific Ambiguity Detection. The Case of German Physics Language

Vitor Lécio Lacerda Fontanella
Hochschule Hannover
Hannover, Germany
vitor-lecio.lacerda-fontanella@hs-hannover.de

Christian Wartena
Hochschule Hannover
Hannover, Germany
christian.wartena@hs-hannover.de

Gunnar Friege
Leibniz Universität Hannover
Hannover, Germany
friege@idmp.uni-hannover.de

Abstract

Many terms used in physics have a different meaning or usage pattern in general language, constituting a learning barrier in physics teaching. The systematic identification of such terms is considered to be useful for science education as well as for terminology extraction. This article compares three methods based on vector semantics and a simple frequency-based baseline for automatically identifying terms used in general language with domain-specific use in physics. For evaluation, we use ambiguity scores from a survey among physicists and data about the number of term senses from Wiktionary. We show that the so-called Vector Initialization method obtains the best results.

1 Introduction

In science, it is common to refer to specific concepts using terms which are also used in everyday language but with a more specific or different meaning. At the same time, terms from science are assimilated into general language, often with a transformed meaning and use. Since these terms have a domain-specific use within science, they are a potential source of ambiguity, generating problems for successful communication and learning.

More specifically, in science education, it has been found that students' conceptions are often related to terms' general meanings non-congruent with the scientific ones (Itza-Ortiz et al., 2003; Clerk and Rutherford, 2000). In physics teaching, for example, words like *work*, *energy*, *momentum*, *impulse*, *power*, and *mass* have a narrower definition and a meaning that often differs entirely from the one used in everyday language (Itza-Ortiz et al., 2003; Song and Carheden, 2014). Song and Carheden (2014) argue that terms with multiple meanings are more difficult to learn, demanding further negotiation, expansion, and correct contextualization of their meanings. They also showed in a study with words from the chemistry teaching (e.g., *solution, polar,* and *compound*) that disassociating the scientific meaning from the one already acquired in everyday life is often hard. Moreover, Itza-Ortiz et al. (2003) show that students' ability to distinguish the different senses of a term correlates with test scores in the corresponding discipline.

By recognizing that terms with different meanings and uses in science and general language represent a learning barrier, their automatic identification within a discipline becomes a relevant task, supporting awareness of their use in teaching (Itza-Ortiz et al., 2003; Strömdahl, 2012; Liu et al., 2022), or even supporting specific teaching strategies for these cases (Vâlcea, 2019).

In Natural Language Processing (NLP), identifying semantic differences between domains (*Synchronic Lexical Semantic Change*) is similar to identifying lexical changes in time (*Diachronic Lexical Semantic Change*). In both cases, we can use properties of word embeddings to detect shifts in the relative positions in the embeddings space. The Synchronic Lexical Semantic Change has recently received attention in engineering requirements (Ferrari and Esuli, 2019; Jain et al., 2019; Mishra and Sharma, 2019) for the detection of potential sources of ambiguity. This task is also investigated in terminology extraction (Hätty et al., 2019), where statistical measures might not identify terms commonly used in specific and general contexts as part of a field's terminology.

Since word embeddings give a concise representation of a word's use (and, according to the distributional hypothesis, the meaning), many authors use them to study the domain-specific meaning of words. However, when computing word embeddings from two different corpora, we will end up with incomparable embedding spaces. The main differences in the proposed approaches deal with the solutions used to overcome this problem.

The present paper aims to compare three methods and a simple baseline for identifying general language words with a deviant meaning in physics. For the evaluation, we use two data sets, one obtained by collecting expert judgments on a small number of nouns and a larger data set derived from Wiktionary. To the best of our knowledge, this is the first evaluation of the automatic identification of general terms with a specific meaning in the science education domain.

2 Related Work

As mentioned above, we cannot immediately compare word embeddings derived from different corpora since the dimensions are randomly initialized, and the same dimensions in the two models will not correspond. The method **Vector Initialization** Kim et al. (2014) was the first to use neural network embeddings and solves this problem by initializing the embeddings' training from the previous corpus embeddings, and then comparing the position of the embedding before and after training. The hypothesis is that the embeddings' displacement after training reflect the word lexical change. This method was initially aimed at diachronic lexical change identification but can also be used for synchronic lexical change. Specific for the synchronic lexical change identification, Ferrari et al. (2017) and Mishra and Sharma (2019) use a variation of the Vector Initialization method to investigate lexical ambiguity between specific domains: they use marked target words before further training the embeddings. However, they need to select target words for the analysis based on their frequencies in both domain-specific corpora.

Other authors proposed to make the vector spaces comparable by defining a linear transformation between embedding spaces based on the solution of the **Orthogonal Procrustes** Problem (Hamilton et al., 2016; Jain et al., 2019; Schlechtweg et al., 2019). In the *Orthogonal Procrustes* analysis, a mapping matrix is determined using *Singular Value Decomposition* that rotates one of the vector spaces. The optimal alignment is the one that minimizes the distances between the embeddings of the same word in both vector spaces. Schlechtweg et al. (2019) added a pre-processing step to the procedure: the alignment of the mean center of the vector spaces before determining the mapping matrix.

The method proposed by Ferrari and Esuli (2019), named here as **Similar Words**, indirectly measures the similarity of embeddings from different spaces: they generate two lists of the most similar words using two vector spaces for a target word. Finally, they compute a rank correlation between the two lists to get an ambiguity score.

The methods above are based on static embeddings, like Word2Vec (Skip-Gram). This way, we obtain a general word representation in each context. Liu et al. (2022) use dynamic embeddings, using the average embedding of up to 1000 BERT embeddings in the corpus. They use a supervised regression model to identify domain-specific terms. Therefore we cannot compare their results to those from the unsupervised approaches discussed above. Martinc et al. (2020) also use averaged contextual embeddings. They use the same fine-tuned BERT embeddings for the general and domain-specific corpus but take the average for examples from each corpus separately. Thus the two averages obtained for each corpus are in the same embedding space and can be compared immediately.

Beyond qualitative evaluation, authors evaluate their method's results using manually created rankings (Ferrari and Esuli, 2019; Schlechtweg et al., 2019; Liu et al., 2022). Ferrari and Esuli (2019) only evaluated his method and ambiguity between specific domains, whereas Schlechtweg et al. (2019) systematically compared methods but, for synchronic lexical change, evaluated the methods only with the ranking of a few words used in general language and in the context of cooking. Liu et al. (2022) evaluated two methods (their own regression model and the unsupervised approach from Martinc et al. (2020)) on three different domains, generating lists of domain-specific terms. These lists were then evaluated manually, using precision as an evaluation measure; however, recall could not be assessed.

3 Experimental setup

In the following, we will present some (technical) information about our implementation of four methods for the identification of terms in general language with a specific sense in physics: **Vector Initialization**, **Orthogonal Procrustes**, **Similar Words** and **Relative Frequency**. The general idea behind the methods we compare was already described in section 2.

We generated three vector spaces using the Skip-Gram method from Gensim (Řehřek and Sojka,

Table 1: Overview of the corpora used.

corpus	Physics	deNews2020
Sentences	796 167	1 000 000
Tokens	15 506 365	17 624 256
Types	252 468	648 959

2011). The first (**model 1**) is obtained from a general language corpus. The second (**model 2**) follows the approach proposed by Kim et al. (2014): we train the model on the physics corpus, but initialize all vectors withe the embeddings from model 1. The third vector space (**model 3**) is obtained from the physics corpus alone. The embeddings have 200 dimensions, and the window size used in training was 5. In all cases we computed embeddings only for words occurring at least 10 times.

For the Vector Initialization method, we calculate the cosine value between the embeddings from model 1 and model 2. Words with the lowest cosine values will assumably have the most significant semantic displacement. For the Orthogonal Procrustes method, we align the vector spaces from models 1 and 3 after the pre-processing step proposed by Schlechtweg et al. (2019). After alignment, we also calculate the cosine values between the embeddings, expecting words with the same usage pattern in the two contexts to be more aligned after the procedure. Finally, for the **Similar Words** method, the ambiguity score will be determined by comparing the most similar words of the terms from models 1 and 3. To evaluate the methods, as a simple baseline, we also sort the words according to the relative frequency, assuming that all words frequently used in physics have a specific meaning in this domain.

4 Data

Identification of terms with domain-specific meaning requires two corpora, a *general language corpus* and a *domain-specific corpus*. The *general language corpus* should be, in principle, non-specific and large, representing, to some extent, everyday language. The domain-specific corpus (*physics corpus*) reflects the communicative context of our interest, namely physics teaching. For the present study, we use a German news corpus *denews2020* (Goldhahn et al., 2012). For the specific corpus, we use a corpus of German texts on physics, mostly high-level textbooks (Lacerda Fontanella et al., 2023). Table 1 gives some details on both corpora.

Table 2: Number of Words in Evaluation. The first column gives the number of words initially collected. The second column the number of the words from this initial collection that occur at least 10 times in both corpora.

	Total	denews ∩ Physics
Survey	48	48
Wiktio. Phy+	766	212
Wiktio. Phy-	135 660	9997

From the literature, we know a few terms that are considered problematic since they have a different meaning in physics than in general language. Such words are e.g., *Arbeit* (work, labor), *Energie* (energy), *Leistung* (power, performance), *Spannung* (tension), *Strom* (electricity, current), *Temperatur* (temperature), *Wärme* (heat, warmth) Strömdahl (2012); Rincke (2010).

However, for a more solid base for evaluation, we collected a set of 48 nouns, including the words mentioned above, along with more problematic and also unproblematic terms. In a survey, we asked participants to what extent, on a scale from 1 (same meaning) to 5 (totally different meaning), the meaning of a term differs in everyday use and the physical context. The ambiguity score for each term is the mean value of the answers in the survey. 14 subjects completed the survey. They were German native speakers with at least a master's degree in physics or physics teaching, including teachers, physicists, and science education researchers. We used survey data for evaluating the methods using the Pearson Correlation between the ambiguity score and the metric obtained for each word from the methods.

For a second experiment, we collected words from Wiktionary and counted how many senses for a word are marked as being specific for physics. E.g., a word like *Kraft* (force) appears with four senses in Wiktionary, one of them referring to physics. Since we compare senses that differ between physics and general language, we evaluated the ranking generated with each method, from potentially more to less ambiguous. We take from Wiktionary the binary information, no physics sense (0), and one or more sense in physics (1). Then, we calculate the area under the curve (AUC) to evaluate the ranking with this binary information. The total of words used (shown in Table 2) is much smaller than the number of words in the Wiktionary, given that the words must appear in both corpora

Table 3: Methods Evaluation.

Method	Survey (Correlation)	Wiktionary (AUC)
Vector Init.	**0.60**	**0.83**
Ortho. Proc.	0.52	0.71
Sim. Words	0.23	0.70
Rel. Freq.	0.58	0.68

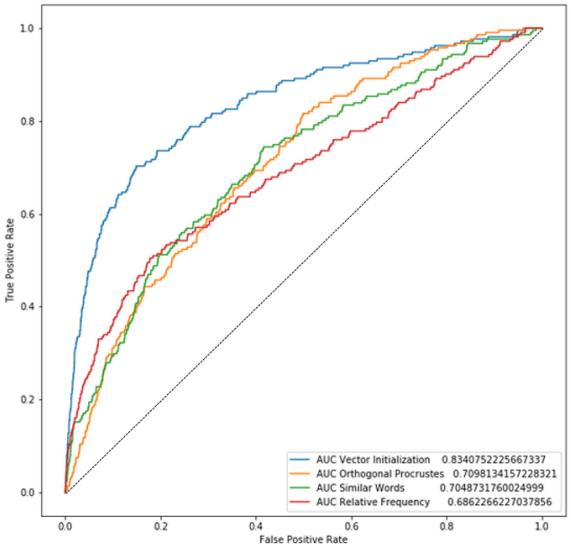

Figure 1: Area under the curve for the methods rankings and the data from Wiktionary.

and hold the minimum frequency requirement for computing an embedding.

5 Results

Table 3 shows the results of the quantitative evaluations of the methods. The method with the best correlation with the survey ambiguity score was the Vector Initialization with a moderate person correlation of 0.60, followed closely by Relative Frequency (0.58). The similar words method performs extremely bad on this task. For the ranking experiment, using much more words from Wiktionary, the Vector Initialization again is the best method, but now this method is clearly much better than the second best method. Relative frequency here is the worst method, though the differences between the other methods are quite small. The good results fron the relative frequency baseline are not surprising, since relative frequency between a specific and general corpus is considered to be an important criterion for terminology identification (Pazienza et al., 2005). However, Vector Initialization clearly outperforms the relative frequency.

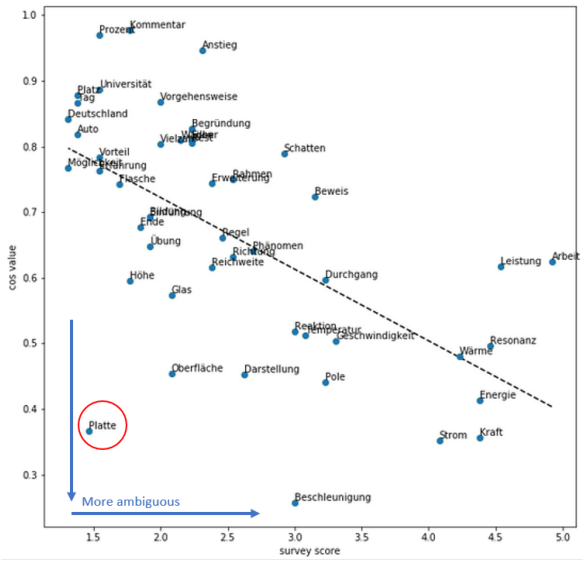

Figure 2: Correlation between **Vector Initialization** cosine values and survey score (Pearson=0.6). The term 'Platte' (board) is the most out the curve.

Finally we look at some qualitative results and examples from the experiments. Table 4 shows the first words in the ranking generated by each method. Here, the terms selected by the Vector Initialization methods make most sense, while especially the Orthogonal Procrustes and Similar Words method give a number of words that seem not to be related to physics at all.

Figure 2 displays the words of the survey with their averaged survey score and computed score. The participants did not perceive *Platte* (board) as ambiguous. Looking at some sentences with the term, we observe that this term is used in German very often referring to music albums. We believe the participants would hardly consider this sense while answering the survey.

Figure 3 shows the effect of the vector initialization method, displaying two terms on their original position and on their position after continued training on the TeCoPhy corpus. We see that the terms move in the direction of other typical physics terms.

6 Conclusion

Lexical ambiguity is a general challenge in communicative situations and an important issue in science education. Identifying domain specific ambiguity is needed to support the appropriate use of language in teaching and specific methodologies for terminology acquisition. In our research, the Vector Initialization method proved to be the most effective for identifying lexically ambiguous words.

Table 4: Twenty top lexical ambiguity candidates.

	Ortho Proc	Vec Init	Similar Words	Rel Freq
1	Heim	Kern	Zwilling	Ladung
2	Neo	Impuls	Ware	Flüssigkeit
3	Aussendung	Masse	Not	Masse
4	Kennzeichen	Winkel	Unterlage	Messung
5	Erhalt	Ladung	Amerikaner	Energie
6	Uniform	Beobachter	Paar	Wärme
7	Toleranz	Flüssigkeit	Lange	Definition
8	Ware	Einheit	Akt	Geschwindigkeit
9	Nerv	Körper	Verbreitung	Eigenschaft
10	Visum	Funktion	Verteilung	Winkel
11	Unterlage	Gas	Schnitt	Intensität
12	Grenzübergang	Feld	Kammer	Theorie
13	Bund	Volumen	Bestimmung	Körper
14	Rausch	Intensität	Zähler	Experiment
15	Hamilton	Ordnung	Produkt	Beschreibung
16	Spaltung	Feder	Ruf	Universum
17	Plus	Spannung	Siemens	Spektrum
18	Weiss	Strom	Signal	Gas
19	Profil	Summe	Brief	Spannung
20	Messe	Dimension	Fluss	Strömung

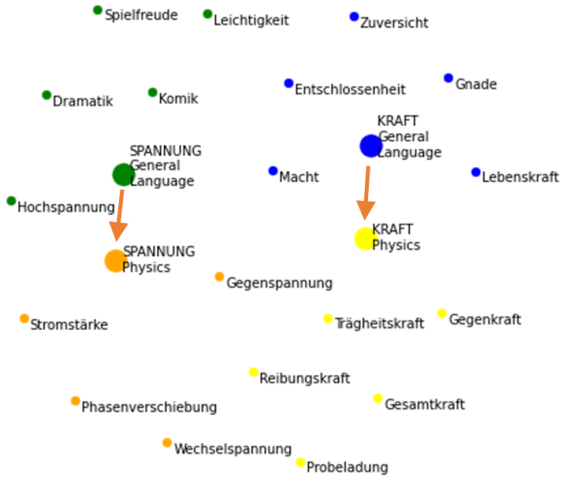

Figure 3: TSNE projection (Maaten and Hinton, 2008) of the most similar word embeddings for the terms Kraft (force) and Spannung (tension, stress, voltage) in model 1 and model 2.

This method achieved the highest Pearson correlation with the survey and AUC calculated with the Wiktionary data. However, in a different study by Schlechtweg et al. (2019), the Vector Initialization method performed poorly when ranking 22 target words. Such conflicting results may be due to the differences in the tasks involved, namely ranking target words versus automatically identifying lexical change.

Moreover, identifying lexical changes can aid in terminology extraction (Hätty et al., 2019), since it can uncover terms with a specialized meaning within a particular domain, despite being frequently used in general language. Such terms may not be found purely based on their frequency. Our research shows that the Vector Initialization method holds more promise than Orthogonal Procrustes as an additional technique for terminology extraction in science education.

A direction for future work is to bring the method to individual occurrences of a word: when finding an instance of an ambiguous word, we would like to be able to see whether the general or domain-specific meaning of the word is intended. This could finally help to see whether students use a word in the correct sense or whether they are misled by the everyday meaning of a specific term. For this purpose, we plan to explore methods based on contextual embeddings and evaluate their applicability to science education.

7 Acknowledgements

This research was funded by the Ministry of Science and Culture of Lower Saxony, Germany, within the PhD program "LernMINT: Data-assisted teaching in the STEAM subjects."

References

Douglas Clerk and Margaret Rutherford. 2000. Language as a confounding variable in the diagnosis of misconceptions. *International Journal of Science Education*, 22(7):703–717.

Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: An nlp approach based on wikipedia crawling and word embeddings. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399.

Alessio Ferrari and Andrea Esuli. 2019. An nlp approach for cross-domain ambiguity detection in requirements engineering. *Automated Software Engineering*, 26(3):559–598.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Anna Hätty, Dominik Schlechtweg, and Sabine Im Schulte Walde. 2019. Surel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on*

*Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Salomon F. Itza-Ortiz, N. Sanjay Rebello, Dean A. Zollman, and Manuel Rodriguez-Achach. 2003. The vocabulary of introductory physics and its implications for learning physics. *The Physics Teacher*, 41(6):330–336.

Vaibhav Jain, Ruchika Malhotra, Sanskar Jain, and Nishant Tanwar. 2019. Cross-domain ambiguity detection using linear transformation of word embedding spaces. *arXiv preprint arXiv:1910.12956*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vitor Lécio Lacerda Fontanella, Tom Bleckmann, Lukas Dieckhoff, Gunnar Friege, and Christian Wartena. 2023. TeCoPhy: A Text Corpus of German Physics Texts. In *Corpus Linguistics in the Digital Era: Genres, Registers and Domains (14th International Conference on Corpus Linguistics)*, pages 122–123, Oviedo, Spain. https://cilc2023.wordpress.com/book-of-abstracts/.

Yang Liu, Alan Medlar, and Dorota Głowacka. 2022. Lexical ambiguity detection in professional discourse. *Information Processing & Management*, 59(5):103000.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Siba Mishra and Arpit Sharma. 2019. On the use of word embeddings for identifying domain specific ambiguities in requirements. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 234–240. IEEE.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In Spiros Sirmakessis, editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 255–279. Springer-Verlag, Berlin/Heidelberg.

Radim Řehřek and Petr Sojka. 2011. Gensim – Statistical Semantics in Python.

Karsten Rincke. 2010. It's rather like learning a language: Development of talk and conceptual understanding in mechanics lessons. *International Journal of Science Education*, 33(2):229–258.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Im Schulte Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Stroudsburg, PA, USA. Association for Computational Linguistics.

Youngjin Song and Shannon Carheden. 2014. Dual meaning vocabulary (dmv) words in learning chemistry. *Chem. Educ. Res. Pract.*, 15(2):128–141.

Helge R. Strömdahl. 2012. On discerning critical elements, relationships and shifts in attaining scientific terms: The challenge of polysemy/homonymy and reference. *Science & Education*, 21(1):55–85.

Cristina Vâlcea. 2019. Teaching technical polysemous words: Strategies and difficulties. In *ICERI2019 Proceedings*, ICERI Proceedings, pages 8388–8394. IATED.