Lexical ambiguity detection in professional discourse

Liu, Yang

2022-09

þÿLiu , Y , Medlar , A & GBowacka , D 2022 , ' Lexical ambiguity detecti
discourse ' , Information Processing and Management , vol. 59 , no. 5 , 103000 . https://doi.org/10.1016/j.ipm.2022.1

http://hdl.handle.net/10138/346636
https://doi.org/10.1016/j.ipm.2022.103000

Information Processing and Management 59 (2022) 103000

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Lexical ambiguity detection in professional discourse

Yang Liu, Alan Medlar, Dorota Głowacka *

University of Helsinki, Finland

ARTICLE INFO

Keywords:
Professional discourse
Specialist terminology
Lexical ambiguity
Word embeddings

ABSTRACT

Professional discourse is the language used by specialists, such as lawyers, doctors and academics, to communicate the knowledge and assumptions associated with their respective fields. Professional discourse can be especially difficult for non-specialists to understand due to the lexical ambiguity of commonplace words that have a different or more specific meaning within a specialist domain. This phenomena also makes it harder for specialists to communicate with the general public because they are similarly unaware of the potential for misunderstandings.

In this article, we present an approach for detecting domain terms with lexical ambiguity versus everyday English. We demonstrate the efficacy of our approach with three case studies in statistics, law and biomedicine. In all case studies, we identify domain terms with a precision@100 greater than 0.9, outperforming the best performing baseline by 18.1–91.7%. Most importantly, we show this ranking is broadly consistent with semantic differences. Our results highlight the difficulties that existing semantic difference methods have in the cross-domain setting, which rank non-domain terms highly due to noise or biases in the data. We additionally show that our approach generalizes to short phrases and investigate its data efficiency by varying the number of labeled examples.

1. Introduction

Professional discourse is the language used by specialists, such as lawyers, doctors and academics, to communicate the knowledge and assumptions associated with specialist training (Kong, 2014). Professional discourse encompasses the spoken, written and visual communication that occurs between specialists (e.g. in academic journals and trade publications) and between specialists and the general public (e.g. between lawyers and their clients). Aside from facilitating communication, professional discourse forms the basis of work as a social practice: it signifies competence, frames professional roles and can impact individuals' identities (Wenger et al., 1998).

Professional discourse can be difficult to understand due to the presence of complex language, and specialist terminology or *jargon* (Links et al., 2019; Schnitzler et al., 2017; Zukswert, Barker, & McDonnell, 2019). Specialist terminology usually takes the form of novel terms and phrases rarely used in everyday language. However, it can also include commonplace words that have a different or more specific meaning in a given professional context, making these terms *lexically ambiguous* (Block, 1986; Cutts, 2015; Ryan, 1985a; Schnitzler et al., 2017). For example, the term *assault* is generally understood to be the act of *inflicting physical harm or unwanted physical contact upon a person*. In the legal context, however, to commit assault is *to cause someone to be put in fear of immediate physical harm*, with the actual *application of physical force* referred to as *battery* (Law, 2015). Similarly, words like *significance* and *independence*, though common in everyday speech, have specialized meanings in statistics (Anderson-Cook, 2010; Kaplan, Fisher, & Rogness, 2009). Lexical ambiguity is considered an important issue in education (Zukswert et al., 2019) and public communication (Cutts, 2015), and has been identified as a problem in numerous professional domains.

* Corresponding author.
E-mail addresses: yang.liu@helsinki.fi (Y. Liu), alan.j.medlar@helsinki.fi (A. Medlar), dorota.glowacka@helsinki.fi (D. Głowacka).

https://doi.org/10.1016/j.ipm.2022.103000
Received 28 February 2022; Received in revised form 20 May 2022; Accepted 16 June 2022
Available online 6 July 2022

For non-specialists, lexical ambiguity is a serious knowledge barrier (Attewell, 1992; Szulanski, 1996). If a user without the necessary expertise searches for medical information or legal advice, for example, and are confronted with unfamiliar terminology, they are at least aware of their lack of understanding. If a domain term is lexically ambiguous, however, and the common usage definition is assumed, then readers may be unaware of their misunderstanding (Block, 1986; Cutts, 2015; Ryan, 1985a; Schnitzler et al., 2017). In a recent example, the book *Outrages: Sex, Censorship and the Criminalization of Love* stated that, contrary to official records, men convicted of sodomy in Britain had been executed later than 1835 (Wolf, 2020). The author, however, had misunderstood the legal term *death recorded* to mean the defendant was executed, when it actually means they were pardoned.[1] Similarly, professionals may fail to communicate information intended for a general audience if they incorrectly assume knowledge of ambiguous specialized terminology (this is sometimes referred to as the *curse of knowledge* Camerer, Loewenstein, & Weber, 1989). For example, over a period of decades, parents in the UK consented to postmortem tissue removal from children without realizing it could result in the harvesting of organs and other body parts because consent forms used the ambiguous term *tissue* (Cutts, 2015). This has clear implications for public communication and the concept of informed consent.

In this article, we focus on the problem of detecting lexically ambiguous terms in professional discourse that could lead to misunderstandings by non-specialists. To solve this problem, we first investigated using semantic shift detection methods (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018), but found them to perform poorly due to unrelated differences in word sense distribution. For example, in the legal domain, *knife* is more commonly used as a verb to describe the act of stabbing someone than to describe a bladed instrument. However, *knife* is not a legal term, its usage is merely biased due to the kinds of situations described in legal proceedings. To address this issue, we developed a semi-supervised approach based on contextual word embeddings (Devlin, Chang, Lee, & Toutanova, 2019) and the Bradley–Terry statistical model for pairwise comparisons (Bradley & Terry, 1952). Our approach assumes the existence of a subject dictionary, such as a legal dictionary (Law, 2015), to provide examples of domain terms, and a comprehensive baseline corpus providing examples of everyday language. We evaluate the performance of our approach in three domains: statistics, law, and biomedicine. We additionally show that our approach generalizes to short phrases and investigate the data efficiency of the method before concluding with a discussion of potential use-cases and limitations.

2. Related work

In this section, we review studies related to lexical ambiguity in legal, scientific and medical discourse, as well as methods for term extraction and semantic shift detection.

2.1. Lexical ambiguity in professional discourse

Research into lexical ambiguity tends to focus on individual domains, leading to inconsistent terminology between fields. Indeed, lexical ambiguity in professional discourse has been referred to as: *paradoxical jargon* (Gowaty, 1982), *multivalent terms* (Ryan, 1985a), *hidden jargon* (Anderson-Cook, 2010), *contextual jargon* (Schnitzler et al., 2017), *non-obvious jargon* (Likwornik, Chin, & Bielinski, 2018) and a type of *language barrier* (Links et al., 2019).

2.1.1. Law

In *The Language of the Law*, Mellinkoff identified many contributory factors to the complexity of legal language, including the *"frequent use of common words with uncommon meanings"*, that render legal documents and legislation inaccessible to laypeople (Mellinkoff, 1963). In common law, the meaning of legal terms is established through precedence, with little regard for their everyday meanings (Charrow, Crandall, & Charrow, 1982). This process has led to pathological examples of meaning change, such as the terms *may* (having the lay meaning of *must*) and *must* (having the lay meaning of *may*) (Charrow et al., 1982). Indeed, there are so many examples of ordinary words with legal meanings that *"the public can be misled, not by language it does not understand, but by language it assumes it does"* (Block, 1986). Furthermore, legal definitions can be so complex, often relying on other legal terms (Block, 1986), that many believe legal terminology should be avoided in communications with the general public altogether (Mellinkoff, 1963; Wydick & Sloan, 2005). In this article, we assume that lexical ambiguity can vary by degree, highlighting which terms are most problematic to non-specialists which could be used to facilitate the communication of legal information to the general public.

2.1.2. Science

Lexically ambiguous terms have been identified as an issue in science communication and education. In sociobiology articles, for example, it was observed that everyday sexual terms carried with them emotional or evocative connotations that might distract from their scientific meaning (Gowaty, 1982), whereas in meteorology there is a need to communicate accurate weather forecasts to the general public (Sivle & Aamodt, 2019). In science education, lexical ambiguity creates problems for non-scientifically inclined students, who assume the general meaning of terms, and for educators, who cannot understand students' confusion (Ryan, 1985a, 1985b). Examples of lexically ambiguous terms have been identified in biology (Rector, Nehm, & Pearl, 2013; Wandersee, 1988; Zukswert et al., 2019), chemistry (Burkholder, 2021; Ryan, 1985a), physics (Taibu, Rudge, & Schuster, 2015; Williams, 1999) and statistics (Anderson-Cook, 2010; Gal, 2002; Kaplan et al., 2009), with the intention of raising awareness among educators. Lexically

[1] This detail was revealed to the author, Naomi Wolf, on the BBC Radio 3 programme Free Thinking https://www.bbc.co.uk/sounds/play/m00057k4 from 21:00–23:00.

ambiguous scientific terms have also been investigated in the context of expert courtroom testimony with the recommendation that problematic language should be avoided to prevent misunderstandings by jurors (Likwornik et al., 2018).

More recently, studies have compared definitions of scientific terms given by scientists and non-scientists (e.g. (Kaplan et al., 2009; Likwornik et al., 2018; Rector et al., 2013; Zukswert et al., 2019)) with each study focusing on a small number of terms identified from the literature. In this article, we present a data-driven approach to rank all terms in the vocabulary by their degree of ambiguity, which could be used to design more comprehensive experimental studies and gain further insights into the impact of lexical ambiguity on education.

2.1.3. Medicine

Lexical ambiguity has become an important topic in effective communication between healthcare professionals and patients (Cutts, 2015). Recently, clinicians' use of medical terminology has been studied in the context of radiation therapy (Schnitzler et al., 2017) and with outpatients from a sleep-disorder clinic (Links et al., 2019) where both studies highlighted the detrimental effect of lexical ambiguity on patient comprehension. Lexical ambiguity also impacts informed consent, with careful consideration needed to avoid using terms that can be unknowingly misunderstood (Cutts, 2015). In this article, we present a method capable of highlighting problematic specialized terms which could be used as a writing aid to enhance communication in clinical settings.

2.2. Semantic shift detection

In natural language processing, the task of identifying terms with different meanings in different contexts is called semantic shift detection (Kutuzov et al., 2018). Semantic shift is the process by which word meaning changes over time, i.e. the meaning of words can narrow (lose meanings), broaden (gain additional meanings), or change due to novel metaphorical or metonymic usage (Thomsen, 2006). Methods for semantic shift detection exploit the distributional properties of word embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to identify terms whose relative positions in the embedding space have changed over time, implying a concordant change in meaning (Kutuzov et al., 2018). Research on semantic shift is usually performed in the context of historical linguistics (Hamilton, Leskovec, & Jurafsky, 2016a; Rodda, Senaldi, & Lenci, 2016). Recently, however, there has been an interest in identifying contemporary linguistic changes in online discourse (Del Tredici, Fernández, & Boleda, 2019; Shoemark, Liza, Nguyen, Hale, & McGillivray, 2019) and within scientific communities (Soni, Lerman, & Eisenstein, 2021). Semantic shift detection methods have been based on non-contextual (Hamilton, Leskovec, & Jurafsky, 2016b) and contextual (Martinc, Kralj Novak, & Pollak, 2020) word embeddings, as well as ensemble approaches (Schlechtweg, McGillivray, Hengchen, Dubossarsky, & Tahmasebi, 2020). The stability of word embeddings can vary significantly across term frequencies (Antoniak & Mimno, 2018; Wendlandt, Kummerfeld, & Mihalcea, 2018), with few methods taking this uncertainty into account (Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015; Liu, Medlar, & Glowacka, 2021).

In this article, we compared our approach with the semantic shift detection methods introduced by Hamilton et al. (2016b) and Martinc et al. (2020) based on Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019), respectively. Semantic shift methods have already been used to identify lexically ambiguous terms in requirements engineering (Ferrari, Donati, & Gnesi, 2017; Ferrari, Esuli, & Gnesi, 2018; Jain, Malhotra, Jain, & Tanwar, 2020). However, as we show in this article, such methods are sensitive to noise and fail to distinguish between domain and non-domain terms.

2.3. Term extraction

Term extraction (also known as glossary extraction) aims to enumerate the terms and phrases associated with a given domain (Damerau, 1993) and has been performed using a wide range of techniques including part-of-speech tagging and statistical methods (Astrakhantsev, Fedorenko, & Turdakov, 2015). Recently, several approaches for term extraction have been built on word embeddings (Hazem, Bouhandi, Boudin, & Daille, 2020), and have been utilized in specific domains, such as medicine (Bay et al., 2021) and cybersecurity (Andrius, 2020).

Term extraction and semantic shift detection are related problems: semantic shift methods have been shown to detect meaning change between domains in benchmark data sets (Schlechtweg, Hätty, Del Tredici, & im Walde, 2019) and semantic shift has been identified as a confounding factor in term extraction (Hätty, Schlechtweg, & im Walde, 2019). Our approach for identifying lexical ambiguity purposely ignores domain terms with the same meaning as everyday English and performs well with low frequency terms rarely included in benchmark data sets.

3. Problem definition and research questions

In this article, we focus on the following problem. Given two corpora: (i) a domain corpus, C_d, and (ii) a non-specialist corpus, C_e, representing baseline English (i.e. everyday English), we want to rank all terms in the shared vocabulary between C_d and C_e, such that the top of the ranking is enriched with domain terms where their meaning differs from common usage.

In many instances, domain-specificity and meaning difference are orthogonal. Domain terms do not necessarily have domain-specific meanings, i.e. their meaning is the same in everyday English and would, therefore, be understood without specialist training. For example, the word *courtroom* is a legal term, but has the same meaning whether used in a legal document or in everyday life. Similarly, there are words that appear to have a domain-specific meaning, but are the result of unrelated biases in the data set. For example, the term *billboard* appears to have a semantic difference because it mostly refers to outdoor display advertising in our

Table 1
Details of corpora and subject dictionaries for each domain.

	Num. documents	Num. tokens	Shared vocabulary size (given term freq. threshold)			Num. labeled examples
			10	100	1000	
Law	55,790	287M	62,756	24,741	7,196	5,781
Statistics	57,398	121M	31,886	11,133	3,011	1,575
Biomedicine	671,446[a]	294M	61,038	22,094	6,799	18,760

[a]Original corpus size, subsampled to 10% of its size in experiments.

law corpus, but refers to the billboard music charts in our baseline English corpus. However, *billboard* is not a legal term and its semantic difference is an artifact of the type of content most likely to be present in each data set. We solve these issues using a semi-supervised approach where subject dictionaries are used to identify examples of domain terms that are, we assume, lexically ambiguous. We explore the following research questions:

RQ1: Can existing semantic shift detection methods be used to extract domain terms whose meaning differs from common usage?
RQ2: Does our approach outperform semantic shift detection methods in terms of precision and coverage?
RQ3: Does our approach generalize to different professional domains?
RQ4: Can our approach detect domain-specific meanings in short phrases?
RQ5: What is the data efficiency of our approach and to what extent does using fewer labeled examples impact performance?

4. Data

We investigated three independent domains as case studies: law, statistics, and biomedicine. For each case study, we needed three data sets: (i) a domain corpus, C_d, (ii) a baseline corpus, C_e, and (iii) a subject dictionary of known domain terms.

Table 1 shows statistics for each domain. We processed subject dictionaries by splitting phrases into individual words, removing duplicate terms and words not found in the shared vocabulary between C_d and C_e. The details of each corpus are as follows:

- **Law:** The law corpus is a collection of high court judgments from England and Wales, including judgments from the UK Supreme Court, the House of Lords and all divisions of the High Court and the Court of Appeal from 1990–2019.[2] High Court judgments are transcripts of spoken court proceedings. We extracted domain terms from the LexPredict legal dictionary.[3]
- **Statistics:** The statistics corpus is composed of articles submitted to ArXiv,[4] and categorized as statistics (i.e. stat.∗ ArXiv categories) for which latex source was available. We could not find a freely available statistics subject dictionary, so created our own word list using the *List of Statistics Articles* Wikipedia page.[5]
- **Biomedicine:** We created a biomedical corpus by downloading the PubMed Central Open Access subset,[6] a digital repository of peer-reviewed biomedical and life sciences scientific literature (Roberts, 2001). For computational reasons, we randomly subsampled the corpus to 10% of its original size. We extracted biomedical terms from an online word list that combined the OpenMedSpel and Raj&Co-Med-Spel-Chek medical dictionaries.[7]
- **Baseline English:** Our approach requires a baseline corpus representing everyday English word usage. For this purpose we used English Wikipedia[8] because of its wide coverage of topics, including the three domains listed above. For each domain, we randomly sampled Wikipedia pages until the number of tokens exceeded that of the domain corpus.

5. Approach

We propose a semi-supervised approach to identify domain-specific terms whose meaning differs from their everyday English usage. We achieve this by combining the Bradley–Terry statistical model with word embeddings generated by BERT and rank all terms by their estimated probability of having a domain-specific meaning.

[2] https://www.bailii.org/databases.html#ew, downloaded June 2019.
[3] https://github.com/LexPredict/lexpredict-legal-dictionary, downloaded June 2019.
[4] https://arxiv.org downloaded June 2019
[5] https://en.wikipedia.org/wiki/List_of_statistics_articles, downloaded June 2019.
[6] https://ftp.ncbi.nlm.nih.gov/pub/pmc/manuscript/, downloaded July 2020.
[7] https://github.com/glutanimate/wordlist-medicalterms-en, downloaded July 2020.
[8] https://en.wikipedia.org, downloaded June 2020.

5.1. The Bradley–Terry model

We used the Bradley–Terry model (Bradley & Terry, 1952) as the basis of our approach. The model has previously been used in information retrieval to produce rankings where there were limited labeled examples for training (Burges et al., 2005; Szummer & Yilmaz, 2011). The Bradley–Terry model is a statistical model to predict the outcome of a tournament between two players, i and j, where $i, j \in \{1, \ldots, K\}$ and the probability that i beats j is:

$$P(i > j) = \alpha_i/(\alpha_i + \alpha_j), \tag{1}$$

where α_i and α_j are positive scores representing the abilities of i and j. This pairwise comparison $i > j$ can represent, for example, i winning a game against j or a preference for i over j, which can be expressed as a logistic regression (Agresti, 2003):

$$logit(P(i > j)) = log\left(\frac{P(i > j)}{1 - P(i > j)}\right) = s_i - s_j, \tag{2}$$

where s_i is a linear predictor: $s_i = log(\alpha_i) = \sum_{r=1}^{p} \beta_r x_{ir}$, of explanatory variables $x_{i1}, \ldots, x_{ip}$ and coefficients $\beta_1, \ldots, \beta_p$ that are estimated via maximum likelihood.

In our approach, the "tournaments" are restricted to two versions of the same word, w: one from the domain corpus, C_d, and one from the baseline English corpus, C_e. These words are represented using the word embeddings w_d and w_e, derived from C_d and C_e, respectively. We consider $w_d > w_e$ to be true if the domain specificity of w_d is higher than w_e, which we assume to be the case if w is in the subject dictionary for a given professional domain and false otherwise. We randomly sample the same number of terms from the shared vocabulary that are not present in the subject dictionary as examples of non-domain terms. We generate the rankings of domain terms by sorting them into descending order of probability.

As we used BERT (Devlin et al., 2019) to generate word embeddings, we applied L1 regularization to the logistic regression model (the dimensionality of BERT embeddings (768) can be higher than the number of known domain terms in some of our experiments). L1 regularization adds an additional penalty term to the log likelihood of logistic regression based on the summation of the absolute values of all coefficients (Tibshirani, 1996). We set the L1 regularization parameter to 5.0 in all experiments, which was found using grid search (data not shown).

5.2. Generating word embeddings

We used the BERT implementation from HuggingFace's Transformers library (Wolf et al., 2020) to generate word embeddings. We used the English BERT-base-uncased pre-trained model with 12 attention layers and a hidden layer size of 768. All hyperparameters were set to their default values. As stated previously, we assume the existence of a domain-specific corpus, C_d, and a baseline English corpus, C_e. We fine-tuned a single BERT model on the union of C_d and C_e using the masked language modeling task for 5 epochs, relying on the contextual nature of BERT embeddings to distinguish between word representations from each corpus (as in Martinc et al., 2020). The ratio of masked to non-masked words was set to 0.15. As this training procedure allows us to generate word embeddings for each corpus using the same model, we do not need to perform any additional post-processing steps to compare embeddings.

We generated word embeddings with our fine-tuned BERT model using the following procedure. We extracted the shared vocabulary of all the terms that appear in both corpora with a term frequency of at least 10. This filters out words that are mutually exclusive to a given corpus and rare terms that are more likely to produce unreliable word embeddings. For each word in the shared vocabulary, we randomly sampled up to 1,000 sentences containing that word from each corpus without replacement and fed them into the fine-tuned BERT model to infer embeddings for each token. We truncated sentences that were longer than 512 tokens (the maximum sequence length for HuggingFace's BERT implementation). We extracted token embeddings by averaging the last four encoder layers (as in Devlin et al., 2019) and merged token embeddings into word embeddings by averaging across tokens. Finally, for each corpus, we averaged the extracted contextual word embeddings for each word to produce domain-specific and baseline English representations, w_d and w_e, respectively.

5.3. Generating phrase embeddings

To highlight the generality of our approach, we show that a model trained on individual words can be used to rank both words and short phrases. To extract phrases, we used the phrase detection method implemented by the Gensim library (Řehůřek & Sojka, 2010) to identify commonly occurring bigrams and trigrams. We allowed n-grams to skip over stop words, making *assault and battery* a valid bigram. The shared vocabulary contained both phrases and their constituent words, i.e. if the vocabulary included the phrase *significance testing*, then we also included the words *significance* and *testing* even if the two words only appeared together.

We represented each phrase by averaging the embeddings of its constituent words. More specifically, each phrase, ph, composed of K words is represented as $ph = (w_1 + w_2 + \cdots + w_K)/K$, following the intuition that phrases containing words with domain meanings are likely to have domain-specific meanings as well.

6. Results

In our evaluation, we assessed domain-specificity using precision@K and whether the meanings of highly ranked terms differed from commonplace usage with estimated semantic shift. We additionally investigated the properties of short phrases and the data efficiency of our approach.

Table 2
Precision for identifying domain terms using Word2vec and BERT-based semantic shift detection methods using term frequency (tf) thresholds 10, 100 and 1000. The best performing method in each set of conditions is bold.

Precision	Model	Statistics			Law			Biomedicine		
		tf = 10	100	1000	10	100	1000	10	100	1000
P@10	Word2vec	0.0	0.0	0.1	0.0	0.1	0.1	**0.3**	0.1	**0.3**
	BERT	0.0	0.0	**0.3**	**0.1**	0.1	**0.2**	0.1	0.1	0.1
P@100	Word2vec	0.02	0.02	0.17	0.01	0.02	0.06	0.16	0.18	0.36
	BERT	**0.03**	**0.09**	**0.24**	**0.06**	**0.19**	**0.36**	**0.17**	**0.3**	**0.39**
P@1000	Word2vec	0.01	0.031	**0.298**	0.018	0.022	0.211	0.146	0.193	0.457
	BERT	**0.046**	**0.14**	0.25	**0.108**	**0.191**	**0.399**	**0.194**	**0.337**	**0.481**

6.1. Baseline selection

We tested the semantic shift detection methods described by Hamilton et al. (2016b) and Martinc et al. (2020) to identify a baseline. Both methods were methodologically similar. In Hamilton et al. Word2vec (Mikolov et al., 2013) is used to train two independent embedding models that are aligned by solving the orthogonal Procrustes problem (Kabsch, 1976). Semantic shift is measured using the cosine distance between embeddings of the same word inferred from different corpora. Whereas, in Martinc et al. word embeddings are extracted from a single fine-tuned BERT (Devlin et al., 2019) model and the cosine distance calculated between average embeddings. We trained Word2vec models with an embedding size of 300 for 5 epochs using the Gensim library (Řehůřek & Sojka, 2010). BERT was trained as described in Section 5.2.

Table 2 shows precision@K for both baselines in all three domains. Precision@K is defined as the proportion of true positives in the top-K words ranked in descending order of cosine distance, where $K = 10$, 100 and 1000. True positives are defined as words found in the corresponding subject dictionary. Word embeddings are known to be unstable at lower term frequencies (Antoniak & Mimno, 2018; Liu et al., 2021; Wendlandt et al., 2018), so we used multiple term frequency thresholds at different orders of magnitude: 10, 100 and 1000. From the perspective of RQ1, in the majority of experiments, the BERT-based approach either outperformed or tied with the Word2vec-based approach. While precision could be as high as 0.48, this required a term frequency threshold of 1000, which excludes a majority of the shared vocabulary (see Table 1). For the remainder of this paper, we will use Martinc et al. (2020) as a baseline, which we refer to as BL(n), where n is the term frequency threshold.

6.2. Performance evaluation

We evaluated the performance of our approach versus the baseline using precision@K (answering RQ2) in three separate domains (answering RQ3). As our approach requires us to randomly sample negative terms to fit the regression model, its performance can vary. We, therefore, created 300 models per domain and present results from the model with the median precision@K.

6.2.1. Manual annotation

As we only have positive examples of domain-specific terms, we manually annotated all words that appear in the top-100 ranked words not present in the subject dictionary for each method in all three domains. We considered a word to have domain-specific meaning if, (i) it is present in an online subject dictionary (e.g. The Free Dictionary[9] contains more comprehensive legal and medical subject dictionaries than are available in machine readable format), (ii) it has a related Wikipedia page that specifically references the domain, or (iii) it is a known disease abbreviation, gene or protein name in the case of biomedicine.

6.2.2. Precision@K

Figs. 1 and 2 show precision@K for rankings that include and exclude the positive examples from subject dictionaries, respectively. For the baseline, BL, we used 10, 100 and 1000 as term frequency thresholds. Our approach used a term frequency threshold of 10.

Our approach consistently ranked domain-specific terms highly, while the performance of the baseline varied depending on the term frequency threshold. In Fig. 1, our approach had a precision@100 greater than 0.9 in all three domains. In statistics, the precision@100 was 91.7% higher than the best performing baseline ($\chi^2(1, N = 200) = 46.095, p = 1.13 \times 10^{-11}$). In law and biomedicine, the precision@100 was 26.0% ($\chi^2(1, N = 200) = 12.502, p = 4.06 \times 10^{-4}$) and 18.1% ($\chi^2(1, N = 200) = 13.085, p = 2.98 \times 10^{-4}$) higher than the best performing baseline. In Fig. 2, we excluded words from the subject dictionaries in evaluation, which reduced the overall precision of all methods across all three domains. While the relative improvement over the best performing baseline in statistics remained high (102.7%, ($\chi^2(1, N = 200) = 29.302, p = 6.19 \times 10^{-8}$)), in law and biomedicine our approach scored only 3.2% and 3.9% higher, neither of which were statistically significant ($\chi^2, p > 0.05$). All other differences in precision@100 between our approach and the baseline were statistically significant ($\chi^2, p < 0.05$) with the exception of BL(100) in biomedicine after filtering out words from the subject dictionary.

[9] https://www.thefreedictionary.com/.

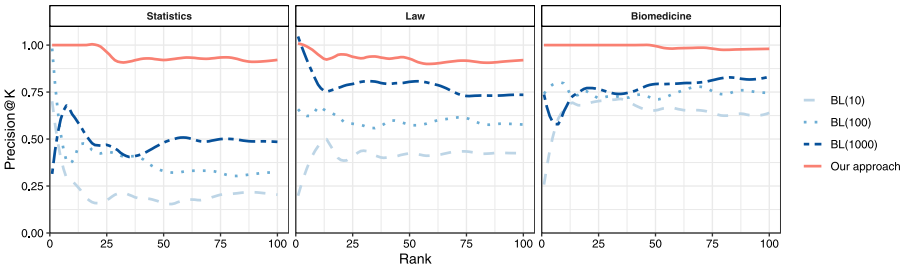

Fig. 1. Precision@K plots for top-100 words in statistics, law and biomedicine. The baseline method (BL) used term frequency thresholds of 10, 100 and 1000. Our approach used a term frequency threshold of 10.

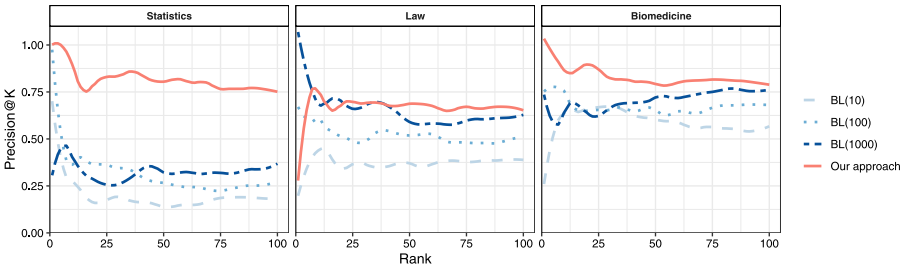

Fig. 2. Precision@K plots for top-100 words in statistics, law and biomedicine after filtering out words from subject dictionaries. The baseline method (BL) used term frequency thresholds of 10, 100 and 1000. Our approach used a term frequency threshold of 10.

Table 3
The top-10 words ranked by each method after filtering out words from subject dictionaries. Domain terms are shown in bold.

Rank	Statistics			Law			Biomedicine		
	Our approach	BL(100)	BL(1000)	Our approach	BL(100)	BL(1000)	Our approach	BL(100)	BL(1000)
1	**completion**	**householder**	sec	vary	**app**	**app**	**mining**	**mining**	**mining**
2	**charting**	**lil**	**directed**	**legislator**	iam	**ac**	**ne**	chapel	chapel
3	**inception**	br	**sir**	**elevation**	ido	**er**	**pearson**	**copa**	**br**
4	**absolutely**	figs	worth	**acta**	**ac**	**rep**	**fe**	**br**	concert
5	**woodbury**	tackles	resort	**deposed**	**er**	frost	**bergmann**	**wta**	customers
6	**conquer**	sect	**inception**	weigh	**rep**	**jr**	**ruby**	**nbl**	**bf**
7	**concentration**	sec	novel	**prospect**	**fours**	assists	theth	**svp**	**ne**
8	**hazards**	**aggregator**	**locality**	**armory**	consonant	**honour**	**draining**	**uss**	**pv**
9	**supervised**	avenue	art	**retiree**	pba	ate	**assembly**	unincorporated	**rev**
10	clayton	**register**	former	**traded**	**bf**	**eat**	**donkey**	touchdown	**accession**

Many of the baseline's top-ranking words are present due to biases in the data or are abbreviations. In Table 3, we show the top-10 ranked terms for each method in all three domains. Domain-specific terms are in bold. In statistics, the baseline ranked many terms related to scientific literature highly, including *figs* (figures), *sec* and *sect* (section) and *art* (from *state of the art*), none of which are domain terms. Similarly, in law and biomedicine, the baseline's top-ranking words were related to legal citations (*app*, *ac*, *er*, *rep*, *jr*) and medical abbreviations (e.g. *uss* (ultrasound scan), *bf* (blood flow), *pv* (pharmacovigilance)). Our approach, however, finds terms that could be misunderstood without specialist knowledge, such as *inception* and *supervised* in statistics (both related to machine learning), and *retiree* in law (which still describes someone who is retired, but has numerous additional legal implications).

6.2.3. Coverage

In Fig. 2, our approach performed similarly in terms of precision to the baseline in law and biomedicine. However, this was only achieved by the baseline after filtering out words with a term frequency of less than 1000, omitting 88.5-90.6% of the shared vocabulary (see Table 1). In the case of law, our approach operated on a vocabulary of 62,756 words, whereas the baseline used only 7,196 words after filtering on the basis of term frequency. Fig. 3 shows the cumulative distribution functions (CDF) over the term frequency for the top-100 ranked words, excluding words found in subject dictionaries. This shows that across domains many of the words filtered out by the baseline are ranked highly by our approach, which includes a wide range of term frequencies among the top-ranked words. A similar plot, including words from subject dictionaries, is included as Supplemental Figure 6 and shows the same trend.

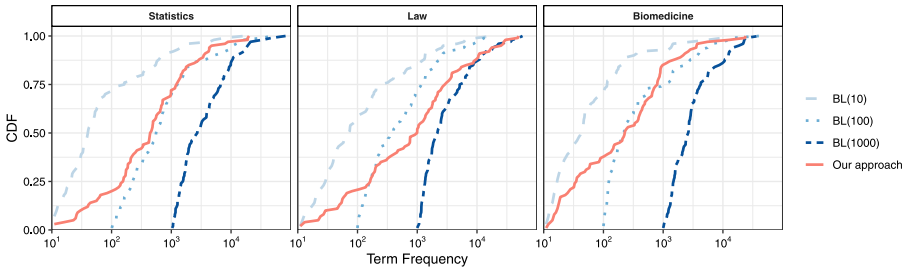

Fig. 3. Cumulative distribution functions (CDF) over term frequency for top-100 words in statistics, law and biomedicine after filtering out words in subject dictionaries.

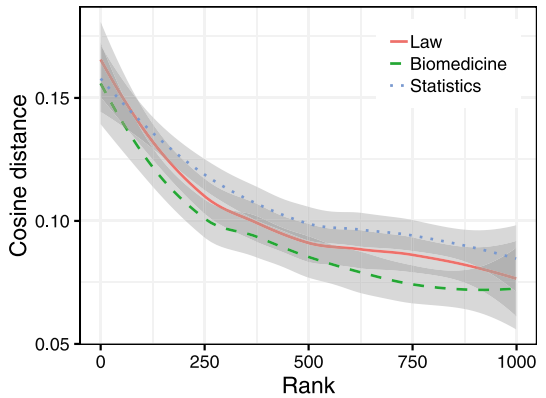

Fig. 4. LOESS curves for the estimated semantic shift of the top-1000 words in statistics, law and biomedicine ranked using our approach.

Table 4
Precision@100 for identifying domain terms using our approach with or without BERT fine-tuning. The best performing method for each domain are shown in bold. The differences in precision are not statistically significant.

	Precision@100		P-value (χ^2)
	With fine-tuning	Without fine-tuning	
Law	**0.65**	0.62	0.659
Statistics	**0.75**	0.7	0.428
Biomedicine	0.79	**0.86**	0.193

6.2.4. Semantic shift

In addition to achieving high precision, our rankings remain broadly consistent with semantic shift. The LOESS curves in Fig. 4 show how the estimated semantic shift decreases with the ranking from our approach. Estimated semantic shift is calculated as the cosine distance from the baseline method. The confidence intervals reflect the fact that non-domain terms with high estimated semantic shift are down-weighted and therefore appear below terms with lower semantic shift in the ranking.

6.2.5. Ablation study

We performed an ablation study to compare the performance of our approach with and without fine-tuning. Table 4 shows the precision@100 in all three domains. In law and statistics, fine-tuning improved precision by 0.03 and 0.05, respectively. In biomedicine, fine-tuning degrades performance by 0.07. However, in all instances, the differences in precision are not statistically significant (see Table 4 for P-values from χ^2 tests). While these results imply that fine-tuning is not strictly necessary to achieve good performance, we cannot guarantee that this holds for all domains and, therefore, recommend the use of fine-tuning.

6.3. Ranking phrases

Our previous results were only concerned with ranking individual terms. To address RQ4, we demonstrate that we can exploit the contextual nature of BERT embeddings to rank arbitrary phrases.

Table 5 shows the top-15 domain-specific phrases in each of the three domains. For each domain, we included where the phrase sits in the overall ranking of all phrases and individual words.

Table 5
The top-15 short phrases and their ranks obtained in the three domains. Domain terms are shown in bold and "∗" indicates terms in the subject dictionary used during training.

Statistics		Law		Biomedicine	
Rank	Phrase	Rank	Phrase	Rank	Phrase
3	second term	22	agricultural purposes	20	**heart attack**
16	third term	57	unexpired term	45	**short tandem repeats**
20	fifth term	123	keel laying	57	**digestive gland**
27	fourth term	167	**de iure**	59	deep purple
32	**large numbers**	185	**charitable trusts**	66	**chest tube**
42	**complete independence**	195	**corrective action**∗	112	**progenitor cell**
58	fully functional	200	eldest child	117	**essential medicines**
145	**functional relationship**	212	**natural habitats**	118	**knockout mouse**
160	**chinese restaurant**	249	growing crops	126	**electron ionization**
168	**nearest neighbour**	258	**chief magistrate**	127	**sodium sulfate**
173	**white noise**∗	299	**circuit judge**	138	ex situ
181	localized version	303	**exclusive distributor**∗	147	**anterior lobe**
190	**one sided**	335	**scheduled monument**	172	upper lobe
191	medical studies	336	**annual pension**	174	**fore limb**
214	**monte carlo**	350	pay tribute	175	**sodium phosphate**

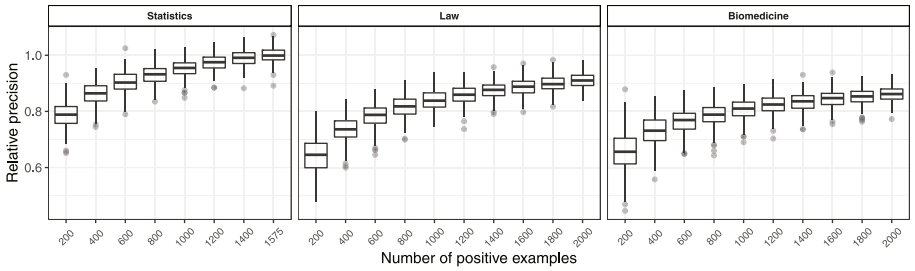

Fig. 5. Boxplots of relative precision in statistics, law, and biomedicine with different numbers of positive examples. In statistics, the precision exceeds 1.0 because our baseline was the model with median performance.

First, we observe that Table 5 has almost no words in common with Table 3 or Supplementary Table 6. Second, there are very few highly ranked phrases. This phenomena can be explained with the phrase *Monte Carlo*: in everyday English the constituent words occur in many different contexts, whereas in statistics they appear together almost exclusively as *Markov Chain Monte Carlo* and *Monte Carlo simulation*. However, the complete phrase *Monte Carlo* appears in relatively fewer contexts in Wikipedia than each word separately (e.g. an administrative area within Monaco, towns of the same name, the Monte Carlo rally). The increased specificity of the phrase, therefore, reduces the degree of semantic shift. This speaks to the robustness of our model as combining two highly ranked words does not automatically result in a highly ranked phrase. Last, phrases in law tend to be ranked lower than the other two domains. This could be partially due to the data coming from spoken legal proceedings, which, even in a courtroom setting, lacks the precision of written prose. It could also be due to legal proceedings being relatively common in people's lives, as demonstrated by their inclusion in biographical articles in Wikipedia.

6.4. Data efficiency

Given that a large subject dictionary may not be available in all domains, we investigated how dictionary size impacts performance (RQ5). We tested dictionaries of increasing size of up to 2,000 words, with the exception of statistics where the complete dictionary contained only 1,575 words. For each dictionary of size, n, we performed 300 experiments. To build the model for each experiment, we randomly sampled n words from the dictionary without replacement as positive examples and n words from the shared vocabulary that did not appear in the dictionary as negative examples.

Fig. 5 shows the relationship between relative precision and dictionary size. The boxplots show considerable variation in precision at lower dictionary sizes, which decreases as the size of the dictionary increases. This is due to the random sampling of positive and negative examples as smaller samples will be more or less representative of the domain than the complete dictionary. We note that the relative precision can be greater than 1.0 because the results presented previously were the median precision@K. Our approach is highly data efficient: we only require 4.0–12.5% of the number of positive examples used to achieve a median relative precision of 0.8 (statistics 12.5%, law 10%, biomedicine 4%) and approximately a third of the number of positive examples to have a median relative precision of 0.9 (statistics 38%, law 33.5%). However, the larger the dictionary, the better the overall performance and the lower the variance.

7. Discussion

In this article, we presented a semi-supervised approach to identify terms and phrases that have the potential to confuse non-specialist readers of professional discourse. We demonstrated the generality of our approach in three separate domains: statistics, law and medicine. Our approach showed that it is possible to identify domain terms (Fig. 2), while producing a ranking that is consistent with differences in meaning (Fig. 4). We were able to achieve these results with $8.7 - 10.6\times$ more words in the shared vocabulary than the best performing baseline (Table 1), the performance of which was highly dependent on term frequency threshold. We highlighted the robustness of our approach by using the same models to incorporate phrases into the ranking, showing that phrases are rarely ranked highly due to their increased specificity compared to their constituent terms (Table 5). Finally, we found that our method has reasonable performance with relatively few positive examples of domain terms, showing that it can be applied in scenarios where a comprehensive subject dictionary is unavailable (Fig. 5).

7.1. Implications

Prior work used semantic shift detection to identify ambiguous domain terms (Ferrari et al., 2017). However, these methods suffer from noise at low term frequencies (Liu et al., 2021), necessitating a high term frequency threshold to maximize precision. Unfortunately, as in our results, such high thresholds can filter out a majority of the vocabulary. Furthermore, semantic shift detection can be misled by term frequency differences between corpora (Table 3), which we avoided by modeling domain specificity. Term extraction methods are unaware of semantic differences (Hätty et al., 2019), making them unsuitable for this purpose.

Our findings have important implications for the design of information retrieval systems. Currently, most systems for specialist information retrieval, such as legal (Maxwell & Schafer, 2008) and medical retrieval (Luo, Tang, Yang, & Wei, 2008), are designed with the implicit assumption that users are themselves specialists and, therefore, able to identify relevant information without additional support (Winkels, Boer, Vredebregt, & van Someren, 2014). Whereas increasing numbers of information seeking studies in similar domains tend to focus on the behavior of non-specialists (e.g. Li, Orrange, Kravitz, & Bell, 2014). We believe that annotating web content on the basis of domain-specificity and meaning difference will highlight gaps in readers' knowledge, foreshadowing potential misunderstandings.

Second, our findings could be utilized in supporting scientific literature search. Users of scientific literature search systems are often interested in finding related literature that goes beyond their immediate area of expertise but, at the same time, they lack sufficient specialist knowledge to identify relevant documents in other research areas. Our approach could be used to help users identify articles with an "intermediate" level of difficulty based on the density of professional discourse. The problem of finding documents of intermediate difficulty has been identified elsewhere, for example, in exploratory search of medical information, where users reported similar issues (Capra, Marchionini, Velasco-Martin, & Muller, 2010). More broadly, the proposed approach could be used in personalization of search results based on users' educational or professional background. For example, adjusting the level of professional discourse in recommended scientific articles to medical students vs. general medical practitioners vs. specialists in specific areas of medicine.

Lastly, from the perspective of specialists writing for a general audience, we believe our method could be used as a kind of readability index, calling attention to text that may be problematic for non-specialist readers (avoiding the kinds of issues highlighted in clinical contexts Cutts, 2015). In this respect, the proposed approach could be used by writers or editors of popular science magazines, blogs or MOOCs to create content that is more accessible to a broader audience.

7.2. Limitations and future work

The work has several limitations. First, we used Wikipedia as a baseline English corpus because it covers all of the domains studied and is written for a general audience. However, the word distribution in Wikipedia is unlikely to match the average English speaker. Moreover, what is considered baseline English varies widely between age groups, education levels and socio-economic backgrounds (Gregory & Carroll, 2018). Thus, depending on the target audience, a more specific baseline corpora might be required to better reflect the linguistic background of targeted users. Second, our approach relies on the presence of a subject dictionary to model domain specificity. While specialist domains often have subject dictionaries, they may not be freely distributed (as in our statistics case study, see Section 4) or even available in machine readable format. Furthermore, even with the presence of a subject dictionary there are numerous potential confounding factors: (i) dictionary terms need to be present in the shared vocabulary to be usable in the Bradley–Terry model, (ii) terms with no semantic difference between specialist and non-specialist domains do not contribute any information to the model, and (iii) vocabularies are constantly evolving and subject dictionaries do not necessarily capture the current state of the professional language usage. In future work, we want to remove the need for getting domain terms from subject dictionaries by, for example, combining our approach with traditional term extraction methods. Finally, while we can use our approach to highlight terms and alert readers to potential comprehension issues, we cannot explain why a given term is highlighted. In statistics, for example, the word *inception* is not obviously related to neural networks for computer vision, which would need to be extracted and summarized from the domain corpus.

CRediT authorship contribution statement

Yang Liu: Methodology, Data curation, Software, Validation, Visualization, Writing – review & editing. **Alan Medlar:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Dorota Głowacka:** Data curation, Writing – review & editing, Supervision, Funding acquisition.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2022.103000.

References

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Anderson-Cook, C. M. (2010). Hidden jargon: Everyday words with meanings specific to statistics. In *Data and context in statistical education: towards an evidence-based society. proceedings of the eighth international conference on teaching statistics*.

Andrius, U. (2020). Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In *Human language technologies–the baltic perspective: proceedings of the ninth international conference baltic HLT 2020, Vol. 328* (p. 39). IOS Press.

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119.

Astrakhantsev, N. A., Fedorenko, D. G., & Turdakov, D. Y. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software*, *41*(6), 336–349.

Attewell, P. (1992). Technology diffusion and organizational learning: The case of business computing. *Organization Science*, *3*(1), 1–19.

Bay, M., Bruneß, D., Herold, M., Schulze, C., Guckert, M., & Minor, M. (2021). Term extraction from medical documents using word embeddings. In *6th IEEE congress on information science and technology* (pp. 328–333). IEEE.

Block, G. (1986). Legal language, lay meanings. *ETC: A Review of General Semantics*, 169–174.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., et al. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on machine learning* (pp. 89–96).

Burkholder, E. (2021). Student and expert conceptions of the word "efficiency". In *IEEE frontiers in education conference* (pp. 1–6). IEEE.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232–1254.

Capra, R., Marchionini, G., Velasco-Martin, J., & Muller, K. (2010). Tools-at-hand and learning in multi-session, collaborative search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 951–960).

Charrow, V. R., Crandall, J. A., & Charrow, R. P. (1982). Characteristics and functions of legal language. In *Sublanguage: studies of language in restricted semantic domains* (pp. 175–189). De Gruyter.

Cutts, M. (2015). Making leaflets clearer for patients. *Medical Writing*, *24*(1), 14–19.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, *29*(4), 433–447.

Del Tredici, M., Fernández, R., & Boleda, G. (2019). Short-term meaning shift: A distributional exploration. In *Proceedings of NAACL-HLT* (pp. 2069–2075).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Ferrari, A., Donati, B., & Gnesi, S. (2017). Detecting domain-specific ambiguities: An NLP approach based on wikipedia crawling and word embeddings. In *Proceedings of the 25th international requirements engineering conference workshops* (pp. 393–399). IEEE.

Ferrari, A., Esuli, A., & Gnesi, S. (2018). Identification of cross-domain ambiguity with language models. In *Proceedings of the 5th international workshop on artificial intelligence for requirements engineering* (pp. 31–38). IEEE.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gowaty, P. A. (1982). Sexual terms in sociobiology: Emotionally evocative and, paradoxically, jargon. *Animal Behaviour*, *30*, 630–631.

Gregory, M., & Carroll, S. (2018). *Language and situation: language varieties and their social contexts*. Routledge.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing, vol. 2016* (p. 2116).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics, Vol. 1* (pp. 1489–1501).

Hätty, A., Schlechtweg, D., & im Walde, S. S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the eighth joint conference on lexical and computational semantics* (pp. 1–8).

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2020). TermEval 2020: TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th international workshop on computational terminology* (pp. 95–100). European Language Resources Association.

Jain, V., Malhotra, R., Jain, S., & Tanwar, N. (2020). Cross-domain ambiguity detection using linear transformation of word embedding spaces. In *CEUR workshop proceedings*: *vol. 2584*, *Proceedings of the third workshop on natural language processing for requirements engineering*.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*(5), 922–923.

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3).

Kong, K. (2014). *Professional discourse*. Cambridge University Press.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web* (pp. 625–635).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1384–1397).

Law, J. (2015). *Oxford dictionary of law*. Oxford: Oxford University Press.

Li, N., Orrange, S., Kravitz, R. L., & Bell, R. A. (2014). Reasons for and predictors of patients' online health information seeking following a medical appointment. *Family Practice*, *31*(5), 550–556.

Likwornik, H., Chin, J., & Bielinski, M. (2018). The diverging dictionaries of science and law. *The International Journal of Evidence & Proof*, *22*(1), 30–44.

Links, A., Callon, W., Wasserman, C., Walsh, J., Beach, M., & Boss, E. (2019). Surgeon use of medical jargon with parents in the outpatient setting. *Patient Education and Counseling*, *102*(6), 1111–1118.

Liu, Y., Medlar, A., & Glowacka, D. (2021). Statistically significant detection of semantic shifts using contextual word embeddings. In *Proceedings of the 2nd workshop on evaluation and comparison of NLP systems* (pp. 104–113).

Luo, G., Tang, C., Yang, H., & Wei, X. (2008). MedSearch: A specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 143–152). ACM.

Martinc, M., Kralj Novak, P., & Pollak, S. (2020). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4811–4819).

Maxwell, K. T., & Schafer, B. (2008). Concept and context in legal information retrieval. In *Proceedings of the 2008 conference on legal knowledge and information systems* (pp. 63–72).

Mellinkoff, D. (1963). *The language of the law*. Wipf and Stock Publishers.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Rector, M. A., Nehm, R. H., & Pearl, D. (2013). Learning the language of evolution: lexical ambiguity and word meaning in student explanations. *Research in Science Education*, *43*(3), 1107–1133.
Roberts, R. J. (2001). PubMed central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*, *98*(2), 381–382.
Rodda, M. A., Senaldi, M. S., & Lenci, A. (2016). Panta Rei: Tracking semantic change with distributional semantics in ancient Greek. In *CLiC-It/EVALITA*.
Ryan, J. N. (1985a). The language gap: Common words with technical meanings. *Journal of Chemical Education*, *62*(12), 1098.
Ryan, J. N. (1985b). The secret language of science or, radicals in the classroom. *The American Biology Teacher*, *47*(2), 91.
Schlechtweg, D., Hätty, A., Del Tredici, M., & im Walde, S. S. (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* (pp. 732–746).
Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1–23).
Schnitzler, L., Smith, S. K., Shepherd, H. L., Shaw, J., Dong, S., Carpenter, D. M., et al. (2017). Communication during radiation therapy education sessions: The role of medical jargon and emotional support in clarifying patient confusion. *Patient Education and Counseling*, *100*(1), 112–120.
Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing*.
Sivle, A. D., & Aamodt, T. (2019). A dialogue-based weather forecast: Adapting language to end-users to improve communication. *Weather*, *74*(12), 436–441.
Soni, S., Lerman, K., & Eisenstein, J. (2021). Follow the leader: Documents on the leading edge of semantic change get more citations. *Journal of the Association for Information Science and Technology*, *72*(4), 478–492.
Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal*, *17*(S2), 27–43.
Szummer, M., & Yilmaz, E. (2011). Semi-supervised learning to rank with preference regularization. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 269–278).
Taibu, R., Rudge, D., & Schuster, D. (2015). Textbook presentations of weight: Conceptual difficulties and language ambiguities. *Physical Review Special Topics-Physics Education Research*, *11*(1), Article 010117.
Thomsen, O. N. (2006). *Competing models of linguistic change: evolution and beyond*. John Benjamins Publishing.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.
Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA.
Wandersee, J. H. (1988). The terminology problem in biology education: A reconnaissance. *The American Biology Teacher*, *50*(2), 97–100.
Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. In *Proceedings of NAACL-HLT*.
Wenger, E., et al. (1998). Communities of practice: Learning as a social system. *Systems Thinker*, *9*(5), 2–3.
Williams, H. T. (1999). Semantics in teaching introductory physics. *American Journal of Physics*, *67*(8), 670–680.
Winkels, R., Boer, A., Vredebregt, B., & van Someren, A. (2014). Towards a legal recommender system. In *Proceedings of the 2014 conference on legal knowledge and information systems, Vol. 271* (pp. 169–178).
Wolf, N. (2020). *Outrages: sex, censorship and the criminalisation of love*. Hachette UK.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).
Wydick, R. C., & Sloan, A. E. (2005). *Plain english for lawyers*. Carolina Academic Press Durham.
Zukswert, J. M., Barker, M. K., & McDonnell, L. (2019). Identifying troublesome jargon in biology: Discrepancies between student performance and perceived understanding. *CBE—Life Sciences Education*, *18*(1), 6.