Natural Language Engineering (2020), **26**, pp. 349–373
doi:10.1017/S1351324919000408

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Using linguistically defined specific details to detect deception across domains

Nikolai Vogler* and Lisa Pearl

Language Science and Cognitive Sciences, University of California, 3151 Social Science Plaza A, Irvine, CA 92697, USA
*Corresponding author. Email: nikolai.vogler@gmail.com

(Received 19 February 2018; revised 7 June 2019; accepted 16 June 2019; first published online 1 August 2019)

Abstract

Current automatic deception detection approaches tend to rely on cues that are based either on specific lexical items or on linguistically abstract features that are not necessarily motivated by the psychology of deception. Notably, while approaches relying on such features can do well when the content domain is similar for training and testing, they suffer when content changes occur. We investigate new linguistically defined features that aim to capture specific details, a psychologically motivated aspect of truthful versus deceptive language that may be diagnostic across content domains. To ascertain the potential utility of these features, we evaluate them on data sets representing a broad sample of deceptive language, including hotel reviews, opinions about emotionally charged topics, and answers to job interview questions. We additionally evaluate these features as part of a deception detection classifier. We find that these linguistically defined specific detail features are most useful for cross-domain deception detection when the training data differ significantly in content from the test data, and particularly benefit classification accuracy on deceptive documents. We discuss implications of our results for general-purpose approaches to deception detection.

Keywords: Deception detection; Cross-domain; Specific details; Linguistic features

1. Introduction

1.1 Automatic deception detection

Deception surfaces in a variety of places, including deceptive opinions about product and service reviews, deceptive statements about specific topics with emotional repercussions, and deceptive representations about personal information, among many others. Notably, detecting deception via the language used to express it is quite difficult, even for humans. This distinguishes text-based deception detection from many other natural language processing tasks, such as speech recognition, parsing, word sense disambiguation, named entity (NE) recognition, relationship extraction, coreference resolution, question answering, sentiment analysis, emotion detection, and nonliteral language use, among many others. In particular, humans without any special training are typically near ceiling performance at these other natural language tasks, while they're often barely above chance at language-based deception detection (Vrij 2000; Newman *et al.* 2003; Ott *et al.* 2011; Ott, Cardie, and Hancock 2013; Fitzpatrick, Bachenko, and Fornaciari 2015).

Moreover, in untrained humans, there's often a pervasive truth bias, where the default assumption is that a statement is truthful (Zuckerman, DePaulo, and Rosenthal 1981; McCornack and Parks 1986; Ott *et al.* 2011; Levine 2014). This is why there's explicit training in the legal and law enforcement domains for exactly how humans can detect deceptive statements more reliably from language alone (Fitzpatrick *et al.* 2015). This is one motivation for applying computational techniques to text-based deception detection—presumably, if the right features are identified, the right algorithm won't suffer from the same bias as humans do.

Natural language engineers have recognized this and begun applying the tools of computational linguistics to automatically detect particular kinds of text-based deception.[a] One type receiving considerable attention is *opinion spam* (e.g., Ott *et al.* 2011; Feng, Banerjee, and Choi 2012; Ott *et al.* 2013; Feng and Hirst 2013; Fornaciari and Poesio 2014; Li *et al.* 2014; Kim *et al.* 2017; Kleinberg *et al.* 2017; Rosso and Cagnina 2017; Narayan, Rout, and Jena 2018). Opinion spam refers to deceptive opinions in service and product reviews that are intended to influence consumer opinion. For example, positive opinion spam may falsely praise an inferior product while negative opinion spam may falsely criticize a superior product. Because reviews significantly impact a potential buyer or user's actions, there's an active market for generating opinion spam. Companies are therefore highly motivated to automatically detect and remove it.

Another type of text-based deception detection receiving some attention concerns false statements that are potentially emotionally charged, such as opinions about certain topics, legally binding testimony, and personal information (Burgoon and Qin 2006; Bachenko, Fitzpatrick, and Schonwetter 2008; Fitzpatrick and Bachenko 2009; Mihalcea and Strapparava 2009; Fornaciari and Poesio 2011; Almela, Valencia-García, and Cantos 2012; Fornaciari and Poesio 2013; Yancheva and Rudzicz 2013; Fitzpatrick *et al.* 2015; Burgoon *et al.* 2016). In these cases, the data may be written opinions (e.g., op-ed pieces) or transcripts of verbal statements (e.g., courtroom testimony and job interviews), and the subjects are highly motivated to succeed at their deception. For example, for all the types of deception mentioned above, the deceivers' reputations may suffer if they're caught. In law enforcement, legal, and employment scenarios, deceivers who are caught can also lose money and employment (or an employment opportunity). In law enforcement and legal scenarios, the deceivers' freedom may also be at risk.

Previous approaches to automatic deception detection have relied on a variety of features, including linguistic features that can be extracted from the text of the document in question. These linguistic features have included *n*-grams (Mihalcea and Strapparava 2009; Ott *et al.* 2011; 2013; Fornaciari and Poesio 2011, 2013, 2014; Fusilier *et al.* 2015; Yu *et al.* 2015), parts of speech (Zhou *et al.* 2004; Ott *et al.* 2011; Fornaciari and Poesio 2011, 2013, 2014; Li *et al.* 2014; Pérez-Rosas and Mihalcea 2015; Yu *et al.* 2015), syntactic structure (Burgoon *et al.* 2003; Zhou *et al.* 2004; Bachenko *et al.* 2008; Feng *et al.* 2012; Feng and Hirst 2013; Yancheva and Rudzicz 2013; Pérez-Rosas and Mihalcea 2015), measures of syntactic complexity (Yancheva and Rudzicz 2013; Pérez-Rosas and Mihalcea 2015), stylometric features (Burgoon *et al.* 2003; Zhou *et al.* 2004; Yoo and Gretzel 2009; Krüger *et al.* 2017), semantically related keyword lists (Burgoon *et al.* 2003; Newman *et al.* 2003; Zhou *et al.* 2004; Mihalcea and Strapparava 2009; Larcker and Zakolyukina 2012; Li *et al.* 2014; Pérez-Rosas and Mihalcea 2015), psychologically motivated keyword lists (Burgoon *et al.* 2003; Newman *et al.* 2003; Zhou *et al.* 2004; Hirschberg *et al.* 2005; Bachenko *et al.* 2008; Yoo and Gretzel 2009; Almela, Valencia-García, and Cantos 2012; Fornaciari and Poesio 2011, 2013; Li *et al.* 2014; Yu *et al.* 2015), sentiment (Yoo and Gretzel 2009; Yu *et al.* 2015), discourse structure (Santos and Li 2010; Rubin and Vashchilko 2012), and NEs (Kleinberg *et al.* 2017), among others.

Notably, automatic deception detection approaches based on these linguistic features tend to do quite well when the content domain is similar for training and testing (e.g., up to *F*-scores of 91.2 on a collection of hotel reviews, using a combination of unigram and syntactic structure features: Feng *et al.* 2012). However, performance noticeably suffers when the training data vary significantly in content from the test data. That is, existing classifiers do well within domain, but struggle to generalize well across domains—they aren't robust to domain change (Krüger *et al.*

[a] We note that the focus has been on detecting texts that were created with the intention to deceive, rather than those that are accidentally misleading due to ignorance or mistaken knowledge on the part of the text's creator. More specifically, these texts might be viewed as doubly deceptive, as they are (i) made up, and (ii) designed to state things the author either doesn't know to be true or in fact knows to be false.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

2017). For example, shifting just the valence of hotel reviews (i.e., training on positive valence reviews but testing on negative valence reviews) can drop the *F*-score to 75 (Ott *et al.* 2013), with deceptive reviews more difficult to detect (70 *F*-score) than true reviews (79 *F*-score). The problem seems exacerbated when the content changes between training and testing. For example, Mihalcea and Strapparava (2009) evaluated their deception detection approach using short narrative essays on emotionally charged topics such as abortion and the death penalty. When the training and test data were from the same content area, *F*-score performance was around 70; when the training and test data were from different content areas, *F*-score performance dropped to around 58–59.

Of course, extensive research into the psychology of deception suggests that a generalized linguistic cue to deception is unlikely to exist (Vrij 2008; Fitzpatrick *et al.* 2015). However, it seems that current automatic deception detection approaches tend to rely on cues that are either (i) very connected to specific lexical items (e.g., *n*-grams and keyword lists), or (ii) linguistically abstract but not necessarily motivated by the psychology of verbal deception (e.g., parts of speech, stylometric features, and syntactic rules). It's therefore useful to explore the impact of using psychologically motivated linguistic cues that are more abstract, as cues with a basis in the psychological theories of how humans generally generate deceptive language may be more applicable across domains (Kleinberg *et al.* 2017). In particular, if these kinds of cues are incorporated into a deception detection algorithm, can they increase cross-domain performance where the training data differ from the test data in terms of valence, content, or other attributes? Recent work in the area of opinion detection in newspaper articles (i.e., whether an article is subjective or objective) suggests that more linguistically abstract features can indeed help this cross-domain challenge (Krüger *et al.* 2017).

1.2 Linguistically defined specific details

With this in mind, we investigate the impact of linguistically defined specific details as features for an automatic deception detection approach. Intuitively, a specific detail is a piece of information that provides a more precise description of the topic. For example, in the domain of hotels, describing a king suite as *a calm, restorative oasis* is more specific than saying *the room was nice*. The amount of detail has been recognized as potentially useful in psychologically motivated approaches for detecting deception and other acts of imagination, including information manipulation theory (McCornack 1992), information management theory (Burgoon *et al.* 1996), criteria-based statement analysis (Steller and Koehnken 1989), and reality monitoring (Johnson and Raye 1981). This is because the amount of detail is thought to correlate with psychological mechanisms underlying the generation of deceptive language, such as the specificity of the memory trace the deceiver is relying on and strategic avoidance of potentially verifiable information. Moreover, several automated approaches have recognized that utterance specificity can distinguish truthful from deceptive texts in certain cases (Burgoon *et al.* 2003; Zhou *et al.* 2004; Burgoon and Qin 2006; Ott *et al.* 2011; Burgoon *et al.* 2016; Kleinberg *et al.* 2017). How these approaches have chosen to implement the idea of specific details in utterances has varied, with the diagnosticity of specific details also varying across different test sets because of the implementation choices. Notably, all these approaches were evaluated within the same domain (i.e., training and testing sets were in the same content area). So, it is unclear how performance generalizes across domain when linguistically defined specific details are among the features a classifier can use.

In this paper, we investigate the utility of a new linguistically defined implementation of specific details for automatic deception detection. We first review previous approaches to automatic deception detection that we base our approach on, focusing on (i) approaches using linguistic features such as *n*-grams that yield the highest performance within domain and (ii) approaches that have incorporated utterance specificity. We then describe the three diverse corpora we use to assess our approach and its cross-domain performance. We then discuss the proposed linguistically defined specific detail features and validate their potential usefulness for detecting deception

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

within these corpora. We subsequently describe our classification scheme, which utilizes these features along with *n*-gram features in a support vector machine (SVM) classifier.

We find that these linguistically defined specific detail features are most useful for cross-domain deception detection when the training data differ significantly in content from the test data. In particular, the more the training data differ in content from the test data in terms of content, the more useful these linguistically defined specific detail features become for detecting deceptive content. Notably, when the content domains from our corpora are maximally different for training and test, using these features alone yields far better performance on detecting deceptive texts compared with true texts. We interpret this to mean that incorporating linguistically defined specific details as features is especially beneficial if it's more valuable to identify potentially deceptive texts (i.e., false negatives are more costly than false positives). We discuss additional implications of our results for general-purpose approaches to deception detection.

2. Related work

2.1 Previous high-performing approaches

2.1.1 Within domain

Of the many linguistic features that have been used, *n*-grams have done surprisingly well on their own within domain, even though small gains can sometimes be achieved when they're combined with other features. For example, Ott *et al.* (2011) found that using unigrams, bigrams, and trigrams with an SVM classifier achieved an 89.0 *F*-score for detecting both true and deceptive hotel reviews with positive valence. Notably, this was only 0.8 behind the best performance (89.8 *F*-score), achieved by using both these *n*-grams and features derived from Linguistic Inquiry and Word Count (LIWC) keyword lists comprised of semantically and psychologically related words (Pennebaker, Booth, and Francis 2007). Feng *et al.* (2012) achieve a 91.2 *F*-score on this same corpus by combining unigrams and bigrams with features involving rules based on syntactic structure. Feng *et al.* (2012) also find that these linguistically abstract structural rules are quite useful on their own, obtaining a 90.4 *F*-score on this same corpus.

As another example within the fake review domain, Fornaciari and Poesio (2014) examined instances of *sock puppetry* in Amazon book reviews, a specific type of positive opinion spam that refers to fake glowing reviews of an individual's own book. Fornaciari and Poesio (2014) found that an approach relying on a subset of *n*-grams and part-of-speech *n*-grams achieved *F*-scores of 78.0 for detecting true reviews and 75.2 for detecting fake reviews.

Mihalcea and Strapparava (2009) constructed a corpus of deceptive language involving essays on three emotionally charged topics: abortion, the death penalty, and opinions about best friends. Using only unigrams, they achieve *F*-scores ranging between 65.9 and 77.0 when training and testing within the same essay content area. On these same data sets, Feng *et al.* (2012) achieve *F*-scores ranging from 67.4 to 81.5 using unigrams and bigrams alone, and increase performance to a range of 71.5–85.0 using a combination of unigrams and features involving syntactic structure rules.

Another example of the utility of syntactically based features comes from Yancheva and Rudzicz (2013), who leveraged a corpus of transcripts from 4- to 7-year-old children. These children were interviewed about a minor transgression involving the child playing with a toy, and in a little over half the transcripts, the child intentionally lied during the interview. Using syntactic complexity features such as average sentence length and mean clauses per utterance, in combination with SVM or Random Forest classifiers, Yancheva and Rudzicz (2013) were able to achieve accuracy scores[b] of up to 91.7.

[b] Note that accuracy (also known as the Rand Index) is a summary statistic similar in spirit to *F*-score in that it provides a single score that penalizes both false positives and false negatives.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

2.1.2 Across domain

Ott *et al.* (2013) conducted one investigation into cross-domain performance that involved changing the valence of hotel reviews—for example, training on positive valence and testing on negative valence, or vice versa. So, the essential content remained constant, but the overall sentiment of that content changed. Using unigrams and bigrams, Ott *et al.* (2013) found a significant drop in *F*-score performance: training on one valence and testing on another yields *F*-scores between 70.3 and 83.0 (compared with training and testing on the same valence, which yields *F*-scores between 85.9 and 89.3). So, even this minimal change of valence causes nontrivial decreases in performance using *n*-gram features.

Another cross-domain investigation involved a data set consisting of hotel, restaurant, and doctor reviews (Li *et al.* 2014). Models trained on hotel reviews were tested on restaurant or doctor reviews, with the idea that restaurant reviews may be more similar in content to hotel reviews than doctor reviews; this is because both hotels and restaurants are places while doctors are service providers. The best *F*-score for testing on restaurant reviews, achieved by a model relying on unigram features, was 78.4; the best *F*-score for testing on doctor reviews, achieved by a model relying on part-of-speech tags, was 67.9. The relative difference in performance accords intuitively with how much the training data differs content-wise from the test data: lower *F*-scores occur for test content that differs more drastically from the training content—even though the general topic of "reviews" remains constant.

The essays data set created by Mihalcea and Strapparava (2009) represents even broader content differences. While opinions about abortion, the death penalty, and best friends fall under the general category of "emotionally charged topics," they vary quite significantly in specific content. Using only unigrams, Mihalcea and Strapparava (2009) trained on two topics' essays and tested on the remaining topic (e.g., training on abortion and death penalty essays, and testing on best friend essays). *F*-score performance using both Naive Bayes and SVM classifiers ranged between 53.6 and 62.0 (and in some cases, cross-domain performance was barely above chance performance of 50.0). Notably, this is far lower than cross-domain evaluations where the content doesn't differ as much between training and test. Feng *et al.* (2012) improve on this performance by using features based on syntactic structure rules, achieving *F*-scores between 66.8 and 70.9. This increased performance highlights the utility of more abstract linguistic features that don't depend on specific keywords, as indicative words might well differ when the content changes.

In a similar vein, Pérez-Rosas and Mihalcea (2015) created and investigated the Open Domain Seception Dataset, which includes sets of seven sentence-long truths and seven sentence-long lies from individual participants. Notably, these sentence-long statements were not restricted by topic, and so cover a broad range (e.g., the participant's age, whether a participant owned multiple Ferraris, what baccalaureate degree a participant had earned, a participant's opinion of the Internet, whether giraffes are taller than zebras, etc.). Moreover, because these sentences occur with no accompanying context (i.e., each individual sentence stands alone rather than all truths/lies being part of a longer statement), this is a particularly difficult cross-domain data set for deception detection. Pérez-Rosas and Mihalcea (2015) found that an SVM classifier relying on unigram features alone achieved an accuracy score of 60.9 (where 50.0 was the random baseline); an SVM classifier relying on parts of speech yielded the highest accuracy score, which was 69.5. So, as with the essays data set, cross-domain deception detection performance increased when more abstract linguistic features were used.

Another data set with potentially broad content changes was that of Bachenko *et al.* (2008), who investigated deception detection on a corpus of individual propositions embedded within real-life high-stakes deceptive narratives: criminal statements, police interrogations, a tobacco lawsuit deposition, and Enron congressional testimony. As with the essays data set, the topics of these propositions could vary quite broadly. Using a set of manually identified deception indicators falling into psychologically motivated categories of *lack of commitment*, *preference for negative expressions*, and *inconsistencies with respect to verb or noun form*, Bachenko *et al.* (2008)

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

used a decision tree to classify propositions as truthful or deceptive. This approach achieved an *F*-score of 70.3 for truthful propositions, and 78.2 for deceptive propositions. These somewhat higher *F*-scores highlight the utility of psychologically motivated features for cross-domain deception detection, though the reliable identification of such features may be more difficult to do automatically.

Investigations by Fornaciari and Poesio (2011, 2013) also focus on propositions within real-life high-stakes narratives coming from Italian court testimonies that involve known calumny or false testimony—this allows their ground truth to be verified. As with the topics of Bachenko *et al.* (2008), it is likely that the topics of the court testimonies vary quite broadly. Using psychologically motivated keyword lists and *n*-grams, Fornaciari and Poesio (2011, 2013) applied different classifiers to corpora of propositions, with the propositions in the training set coming from different court cases than the propositions in the test set. This approach achieved an *F*-score of 79.6 for true propositions (using logistic regression), but an *F*-score of only 63.0 for false propositions (using an SVM). This highlights the difficulty in catching deceptive statements, an area that may be useful for cross-domain approaches to focus on.

2.2 Previous approaches using specific details

2.2.1 Within domain

Burgoon *et al.* (2003) investigated whether utterance specificity, as implemented by the rate of adjectives and adverbs, could be used to discriminate between truthful and deceptive interviews about a mock theft. They didn't find that this feature was useful (i.e., there were nonsignificant differences between truthful and deceptive interviews).

Zhou *et al.* (2004) investigated the utility of utterance specificity, implemented as either *spatiotemporal information* or *perceptual information*. Spatiotemporal information involved locations of people or objects, when an event happened, or explicit descriptions of a sequence of events; perceptual information involved sensory experiences, such as sounds, smells, sensations, and visual details.[c] These were among the potential features a classifier could rely on to discriminate truthful e-mails from deceptive e-mails about items to be salvaged from an overturned jeep in a desert. Both information types emerged as highly weighted features used by decision tree and neural network classifiers. This contrasts with the findings of Burgoon and Qin (2006), who investigated the utility of features based on perceptual information (implemented as the number of sensory details and the ratio of sensory terms to total terms) for discriminating between truthful and deceptive answers to interview questions. Neither perceptual feature's use was significantly different between truthful and deceptive answers. Burgoon and Qin (2006) also investigated a linguistic definition of utterance specificity, based on the number of modifiers used (e.g., the adjectives *calm* and *restorative* in *a calm, restorative oasis* would be modifiers). Modifier use also didn't differ significantly between truthful and deceptive answers.

Burgoon *et al.* (2016) investigated a different linguistic implementation of utterance specificity on a more naturalistic, high-stakes data set involving conference calls about financial reports. Here, utterance specificity was determined by a composite of spatiotemporal keyword terms, determiners, cardinal numbers, subordinating conjunctions, prepositions, and adjectives. With this implementation, Burgoon *et al.* (2016) discovered that deceptive statements had higher specificity than truthful statements and, in fact, that almost all individual specificity features were higher for deceptive versus truthful statements. This suggests that this version of specificity (which includes several linguistically abstract features) has diagnostic potential, though it has not yet been incorporated into a classifier.

[c] We note that there was no description about how these details were automatically extracted, and so it's unclear if they were based on linguistic constructions, keyword lists, or something else.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

In a similar vein, Kleinberg *et al.* (2017) investigated how NEs could be used as a heuristic for specific details, with the idea that NEs often encode more specific information than equivalent non-NEs (e.g., *Michigan Avenue* is more specific than *the street*). Kleinberg *et al.* (2017) examined the proportion of NEs in truthful versus deceptive hotel reviews of both positive and negative valence, finding that truthful reviews of both valences had a higher NE proportion than deceptive reviews. As with the findings of Burgoon *et al.* (2016), these results suggest that this NE-based version of specificity has diagnostic potential, though it has not yet been incorporated into a classifier.

Another approach to more abstract implementations of specific details involves automatically inferring the keywords and phrases that capture specific details within a content domain. In the domain of positive valence hotel reviews, Feng and Hirst (2013) introduce the idea of review profiles that incorporate different "aspects" of content, both *distinct* aspects that typically correspond to proper names (e.g., *Michigan Avenue*) and *general* aspects that reflect topics associated with hotels (e.g., *breakfast*). Both types of aspects are automatically inferred by hierarchical agglomerative clustering over noun phrases in the hotel reviews, with a preset threshold for when to stop clustering based on average cluster distance. Hotel review profiles were then constructed from collections of these aspects and the aspects' relative frequency within a reference set of truthful reviews.[d] When profile compatibility features are included with unigram, bigram, and syntactic structure features, an SVM-based classifier is able to achieve an *F*-score of 91.4 (a modest gain over the version without these compatibility features, which achieved an *F*-score of 90.2).

2.2.2 Across domain

To our knowledge, we are the first to investigate the use of specific details when detecting deception across domains. However, Li *et al.* (2014) note that specific details are, by nature, specific to a particular content domain. So, in this vein, certain LIWC topics could be viewed as representative of specific details within a particular domain. For example, the LIWC *space* topic may correspond to spatial specific details. Li *et al.* (2014) therefore investigated which LIWC topic features were useful when detecting deception across the domain of product and service reviews (i.e., training on hotel reviews and testing on either restaurant or doctor reviews). They found that the *space* and *home* topics were useful for detecting deception in hotel reviews, the *ingest* topic was useful for restaurant reviews, and the *health* and *body* topics were useful for doctor reviews.

This motivates our current exploration into the utility of linguistically defined specific details in a classifier meant to detect deception across domains. It could be that there is some commonality in the linguistic form of specific details that can be harnessed across domains, irrespective of the content of these details. Our approach to linguistically defining specific details, described in Section 3.2.1, follows this intuition.

3. Approach

3.1 Corpora

We investigate deceptive content occurring in three separate data sets, which represent a broad sampling across the domains of online hotel reviews (Ott *et al.* 2011, 2013), essays on emotionally charged topics (Mihalcea and Strapparava 2009), and personal interview questions (Burgoon *et al.* 1999; Burgoon and Qin 2006). We note that all three corpora were generated by telling participants to intentionally deceive, and so represent intentional (rather than unintentional) deception. Notably, the corpora include both noninteractive and interactive venues (Fitzpatrick *et al.* 2015) as shown in Table 1, and so also represent diversity in manner of deception elicitation.

[d] In Section 3.2.2, we describe a feature scaling method that intends to exploit a similar notion of compatibility against a reference distribution, applied to individual features.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 1. Examples of noninteractive and interactive deception types from Fitzpatrick *et al.* (2015), with the types bolded that are covered by our three corpora

Noninteractive	Interactive
Online hotel reviews	**Interviews**
Opinion essays	Court testimony
Voicemail	Phone conversations
Political speeches	Chat/e-mail

Table 2. Details of the three corpora, which include differing degrees of content variation. Number of (T)ruthful and (D)eceptive text samples are shown, along with average text length in tokens

		#		Avg length	
Corpus	Valence/topic	T	D	T	D
DOS	+	400	400	140	129
	−	400	400	204	198
Ess	Abortion	100	100	111	84
	Death Penalty	98	98	112	92
	Best Friend	98	98	99	76
DI	Twelve questions	360	360	168	147

DOS, Deceptive Opinion Spam; Ess, Essays; DI, Deceptive Interview.

More specifically, the online hotel reviews and essays represent deception elicited in a noninteractive manner (in these cases, through Amazon's Mechanical Turk interface), which means they don't necessarily require responses in real time. So, deceivers can create their deceptive content without time pressure, whether the content is a review about a service or product (online hotel reviews) or an introspective opinion about a personal belief (opinion essays). In contrast, the interview responses are deception elicited in an interactive, conversational environment and so require prompt responses. This means the deceiver must create content quickly in the moment, without time to reflect or edit later on.

Additionally, the three corpora we use (shown in Table 2) vary significantly with respect to content, from very narrow differences to very broad differences. The online hotel reviews in the Deceptive Opinion Spam (DOS) corpus (Ott *et al.* 2013) are used as a benchmark in automatic deception detection approaches (Feng *et al.* 2012; Feng and Hirst 2013; Li *et al.* 2014; Kleinberg *et al.* 2017) and differ only in the sentiment (positive valence vs. negative valence). So, they represent a minor difference in content. The essays about personal beliefs in the Essays (Ess) corpus (Mihalcea and Strapparava 2009) are on the topics of abortion, the death penalty, and traits of best friends. Thus, the content of each essay subset differs fairly substantially, though all are part of the general category of "emotionally charged topics." The Ess data set has also been used as a benchmark for previous deception detection studies (Feng *et al.* 2012) because of this markedly broader range of content. The interviews in the Deceptive Interview (DI) corpus (Burgoon *et al.* 1999) involve 12 separate questions for each individual, and the question topics range from personal background to issues of morality. So, this corpus represents the broadest content variation. We describe each corpus in more detail below.

3.1.1 Deceptive Opinion Spam

Ott *et al.* (2011) created the first larger-scale corpus of 400 truthful and 400 deceptive positive sentiment reviews from TripAdvisor. These reviews were of the 20 most popular Chicago-area hotels. Positive truthful reviews were mined from 5-star reviews on TripAdvisor.[e] Positive deceptive reviews were commissioned from Amazon's Mechanical Turk and manually checked for both correctness and plagiarism (using http://plagiarisma.net). For the task, Turkers were presented with the name and website of a hotel and given 30 minutes to write a fake positive review impersonating a customer, as if the Turkers were members of the hotel's marketing department. Turkers had to be located in the United States, have a 90%+ approval rating, and were only allowed to submit once. Additionally, Turkers were informed that reviews must "sound realistic and portray the hotel in a positive light" and be of "sufficient quality" (i.e., they can't be plagiarized, too brief (<150 characters), written for the wrong hotel, or unintelligible).

Later, Ott *et al.* (2013) added 400 truthful and 400 deceptive negative sentiment reviews with the same method for the same hotels in order to facilitate research on the interaction between sentiment and deception. Negative truthful reviews were mined from 1- or 2-star reviews on Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp. Negative deceptive reviews were again commissioned from Amazon's Mechanical Turk and manually checked for correctness. Table A1 in Appendix A shows samples of positive truthful and deceptive reviews.

To obtain a human detection benchmark, Ott *et al.* (2011) asked three undergraduate student volunteers to annotate one fold of their cross-validation experiment (160 opinions). The three judges obtained an average accuracy of 57.3, with an F-score of 66.1 for truthful reviews (precision: 54.7, recall: 84.2) and an F-score of 40.9 on deceptive reviews (precision: 69.3, recall: 30.5). These results accord with the known human *truth bias* (Vrij 2008), where judges are prone to overclassify texts as truthful when attempting to detect deception. Ott *et al.* (2011) note that the truth bias was barely diminished even after informing the judges of the truthful versus deceptive base rate in the data set ahead of time and repeating the study.

3.1.2 Essays

Mihalcea and Strapparava (2009) created the Ess corpus, which contains labeled true and deceptive opinions on abortion, the death penalty, and feelings about a best friend. This data set was also created by using Amazon's Mechanical Turk. Each essay topic contains approximately 100 truthful and 100 deceptive essays containing statements like those in Table A2 in Appendix A. For the abortion and death penalty topics, the guidelines for the contributors were to write at least four to five sentences about their true opinion on the topic as if they were preparing a speech for a debate, and then repeat the same process for their deceptive opinion. For the best friend topic, contributors were asked to write truthfully about the reasons they like their actual best friend, and deceptively about the reasons they like a person they can't stand.

3.1.3 Deceptive Interview

Burgoon *et al.* (1999) created the DI corpus by transcribing 122 verbal interview records where participants answered 12 interview questions, alternating between truthful and deceptive responses. Alternations occurred after every three questions with some participants beginning with a truthful block of questions and others beginning with a deceptive block. Questions range anywhere from open-ended ethics (*If you found a wallet containing $1000 and no identification, what would you do with it? Why?*) to personal background (*Please describe your current or last occupation.*) to personal views (*What political issues do you feel strongly about and why do you*

[e] Notably, the truthfulness of mined reviews can't be verified and so isn't considered gold standard. However, subsequent research has found that the deception rate on online review portals tends to be small (Ott, Cardie, and Hancock 2012; Mayzlin, Dover, and Chevalier 2014).

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 3. Linguistically defined specific detail features, including a description and relevant example

Feature	Description	Example
PP modifier: count	# PP modifiers	of cool, uncluttered comfort
PP modifier: length	# words in PP modifiers	*of cool, uncluttered comfort* = 4
AdjP modifier: count	# AdjP modifiers	cool, uncluttered
AdjP modifier: length	# words in AdjP modifiers	*cool, uncluttered* = 2
Numbers: count	# numbers	two
Proper nouns: count	# proper nouns	Navy Pier
Consecutive nouns: count	# nouns occurring next to each other	airport shuttle

PP, prepositional phrase; AdjP, adjective phrase.

think you feel strongly about them?)—see Table A3 in Appendix A for a complete list of interview questions. Table A4 in Appendix A includes a sample true answer and a sample deceptive answer to an interview question about the participant's educational background.

3.2 Features

3.2.1 Feature definition

To identify potential linguistic definitions for specific details, we first manually examined several samples of text from the three corpora and manually extracted examples of what seemed to us to be specific details. An example of this manual identification procedure is shown in Table A5 in Appendix B. Upon inspection of this sample of specific detail examples, we discovered several recurring linguistic patterns. These included exact number words, proper nouns, and modifiers in the forms of adjectives, nouns, and prepositions. In each case, it seems that these linguistic forms are a way to make the meaning more precise. For example, exact number words give a precise quantity, proper nouns indicate a specific referent in the world, and modifiers to a noun – whether adjectives, prepositions, or other nouns – identify a more precise type of that noun. We note that some of these cues accord with Kleinberg *et al.* (2017), whose NE-based approach identified that the use of cardinal numbers and certain proper nouns differed between truthful and deceptive hotel reviews in the DOS corpus.

With this in mind, we created seven linguistically defined specific detail features, shown in Table 3. These features are based on the following: prepositional phrase (PP) modifiers (e.g., *of cool, uncluttered comfort* in *a haven of cool, uncluttered comfort*), adjective phrase (AdjP) modifiers (e.g., *cool, uncluttered* in *cool, uncluttered comfort*), exact number words (e.g., *two* in *two minutes*), proper nouns (e.g., *North America*), and noun modifiers that appeared as consecutive noun sequences (e.g., *airport shuttle*). For all features, we used the total normalized number (*count*). For phrasal modifiers (PP and AdjP), we additionally included their average normalized length (*length*).

Of course, not all specific details we manually identified are captured via the linguistic definitions we propose and not all the information captured by these linguistic definitions are necessarily specific details. Still, after defining these linguistic features, we used the Stanford parser (Klein and Manning 2003) to automatically generate constituency parse trees that allowed us to then automatically extract these features; we then performed a simple analysis of the features' diagnosticity on texts within each corpus and across all three corpora together (Table 4). In particular, we calculated their mean values and standard deviations within truthful texts and within deceptive texts. While classifiers are capable of capturing much subtler patterns and using those

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 4. Specific detail feature means (with standard deviations in parentheses) in (T)ruthful versus (D)eceptive texts within three corpora and across all three corpora aggregated together. Results from a two-sample, independent *t* test are shown, with statistically significant results ($p \leq 0.05$) bolded

	Opinion Spam		Essays		Interview		All three	
Feature	T	D	T	D	T	D	T	D
PP count	1.5	1.4	1.4	1.2	1.0	0.9	1.36	1.26
	(1.0)	(0.6)	(0.8)	(0.8)	(0.7)	(0.7)	(0.9)	(0.7)
	$p =$ 0.126		$p =$ **0.001**		$p =$ 0.184		$p =$ **0.002**	
PP length	1.4	1.3	1.3	1.2	1.2	1.2	1.31	1.26
	(0.4)	(0.3)	(0.3)	(0.4)	(0.5)	(0.4)	(0.4)	(0.35)
	$p =$ **0.008**		$p =$ **0.009**		$p =$ 0.384		$p =$ **0.0007**	
AdjP count	0.2	0.1	0.1	0.1	0.1	0.1	0.15	0.12
	(0.3)	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.25)	(0.18)
	$p =$ **0.0003**		$p =$ 0.16		$p =$ 0.162		$p =$ **6.3e-5**	
AdjP length	1.7	1.4	1.2	1.0	0.9	0.8	1.41	1.19
	(1.8)	(1.8)	(1.8)	(1.8)	(1.4)	(1.4)	(1.74)	(1.71)
	$p =$ **0.001**		$p =$ 0.353		$p =$ 0.271		$p =$ **0.0007**	
Numbers	0.2	0.1	0.1	0.0	0.1	0.1	0.19	0.11
	(0.3)	(0.2)	(0.2)	(0.1)	(0.2)	(0.2)	(0.25)	(0.16)
	$p =$ **3.1e-23**		$p =$ **3.4e-7**		$p =$ 0.07		$p =$ **2.0e-24**	
Proper nouns	0.5	0.5	0.1	0.2	0.2	0.2	0.36	0.37
	(0.6)	(0.5)	(0.4)	(0.3)	(0.3)	(0.4)	(0.53)	(0.45)
	$p =$ 0.73		$p =$ 0.45		$p =$ 0.80		$p =$ 0.55	
Consecutive nouns	0.5	0.6	0.3	0.2	0.2	0.1	0.39	0.39
	(0.5)	(0.4)	(0.5)	(0.4)	(0.3)	(0.2)	(0.46)	(0.41)
	$p =$ 0.059		$p =$ 0.07		$p =$ **0.048**		$p =$ 0.97	

PP, prepositional phrase; AdjP, adjective phrase.

patterns to classify individual documents, this serves as a first pass look at these features' potential ability to detect deception.

Within each corpus, we find that at least one feature is already diagnostic of truthful versus deceptive texts (DOS: PP length, AdjP count, AdjP length, Numbers; Ess: PP count, PP length, Numbers; DI: Consecutive nouns), and two have several diagnostic features (DOS and Ess). Interestingly, in every single case, the mean of the feature value in the deceptive texts is less than that of the truthful texts. This accords with the idea that deceptive texts use fewer specific details, and suggests these features can capture that linguistic behavior to some extent within domain. So, we have reason to believe a classifier relying on these features may succeed at detecting deception. Across corpora (where all truthful texts were combined into a single truthful corpus and all deceptive texts were combined into a single deceptive corpus), we see a similar pattern: five of the seven features have a significantly lower mean in deceptive texts, when compared to truthful texts (PP count, PP length, AdjP count, AdjP length, Numbers). This suggests these features may also be useful for detecting deception not just within domain but also across domain. With this in mind, we investigate their utility within a deception detection classifier.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

3.2.2 Feature scaling

To increase the prominence of potentially distinctive feature values, we apply a scaling method to the raw feature values. This scaling method was introduced by Pearl and Steyvers (2012) and so we call it *PS Scaling*. PS Scaling is a preprocessing method previously used to successfully detect authorship deception (Pearl and Steyvers 2012; Pearl, Lu, and Haghighi 2016), and accentuates feature values that are diagnostic of a particular class.

Each feature f is scaled separately by first log-transforming all values for f. This creates a distribution of values that are approximately normally distributed, given a large enough sample size. Here, we estimate two normal distributions for each f: one for truthful (T) texts and one for deceptive (D) texts. Then, for any individual text's logged feature value f_v, we compute the log-odds ratio as in (1):

$$\log\left(\frac{p\left(f_v \mid T\right)}{p\left(f_v \mid D\right)}\right) \tag{1}$$

This yields a positive value for features that are more indicative of truthfulness, because the likelihood of the feature value from the Truthful distribution is higher than the likelihood of the feature value from the Deceptive distribution. Similarly, this yields a negative value for features that are more indicative of deception. This scaling process also yields a value near 0 for feature values whose likelihoods are similar across the Truthful and Deceptive distributions.

3.3 Classification scheme

Similar to prior deception detection approaches (Mihalcea and Strapparava 2009; Ott *et al.* 2013; Feng *et al.* 2012; Feng and Hirst 2013), we perform our experiments with an SVM classifier using a linear kernel.[f] Linear SVM classifiers learn a max-margin linear hyperplane in a high-dimensional feature space in order to separate two classes of data. This has the effect of automatically determining which features are relevant for classification, as a function of the weights they're assigned. So, any feature receiving a nonzero weight is useful for classification. This allows us to inspect which features the classifier viewed as most diagnostic. A featurized document **x** is thus classified using learned weights **w** and bias b.

$$\hat{y} = sign(\mathbf{w} \cdot \mathbf{x} + b) \tag{2}$$

For each experiment, the SVM's regularization parameter C, which controls the size of the margin between support vectors, is tuned using nested cross-validation[g] based on the F-score on the training data.

Because some of the comparison SVM classifiers described below will rely on n-gram features, we also extract unigrams and bigrams that occur more than once and transform them into TF-IDF values. For a description of the specific detail features also used in classification, see Section 3.2.1.

4. Results

We begin by setting up deception detection scenarios that are within the same domain (*within domain*), differ narrowly by domain (*narrow change*), or differ more broadly by domain (*broad change*). Deception detection within domain occurs by training on the data from the same domain that will be used for testing (e.g., training on both positive and negative valence opinion spam, and testing on both positive and negative valence opinion spam). This is done for each of our

[f] We use the scikit-learn implementation, which is based on libsvm.

[g] Our parameter grid consists of the values [0.001, .01, .1, 1, 10, 100, 250, 500, 750].

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 5. Corpus splits used for investigating different degrees of cross-domain deception detection within the DOS positive (DOS+) and DOS negative (DOS−) valence corpora, the Ess corpora about abortion (Ess-Ab), best friends (Ess-BF), and the death penalty (Ess-DP), and the DI corpus. We report the JSD to estimate the domain similarity of the narrow and broad cross-domain experiments

	Train data	Train size	Test data	Test size	JSD
Within domain	DOS	640	Cross-validation	160	—
	Ess	590–595	Cross-validation	118–9	—
	DI	576	Cross-validation	144	—
Narrow change	DOS+	400	DOS−	400	29.9
	DOS−	400	DOS+	400	29.9
	Ess-Ab+BF	396	Ess-DP	196	29.6
	Ess-Ab+DP	396	Ess-BF	196	34.1
	Ess-BF+DP	392	Ess-Ab	200	36.2
Broad change	DOS+Ess	1392	DI	720	38.7
	Ess+DI	1312	DOS	800	37.9
	DOS+DI	1520	Ess	592	37.0

DOS, Deceptive Opinion Spam; Ess, Essays; DI, Deceptive Interview; JSD, Jensen–Shannon divergence.

three corpora: DOS, Ess, and DI. These experiments can be compared directly with prior work on within-domain deception detection.

Narrow changes to domain content are implemented by training on delineated subsets within the DOS and Ess data sets and testing on other subsets. For the DOS corpus, training occurs on texts with one valence (e.g., positive) and testing occurs on texts with the other valence (e.g., negative). For the Ess corpus, training occurs on essays from two of three emotionally charged topics (e.g., Abortion and Best Friends) and testing occurs on essays from the third emotionally charged topic (e.g., Death Penalty).

Broad changes to domain content are implemented by training on data from two of the corpora (e.g., DOS and Ess) and testing on the data from the third corpus (e.g., DI). Because the content differs more strikingly across these corpora, this serves as the strongest test case of cross-domain deception detection.

In Table 5, we outline both the training and testing sizes and the splits for each domain change scenario. Additionally, we include the Jensen–Shannon divergence (JSD) of the training and test sets' unigram probability distributions for narrow and broad change scenarios. The JSD is a smoothed, symmetric version of the Kullback–Leibler divergence commonly used as a domain similarity metric in cross-domain sentiment analysis (Remus 2012; Ruder, Ghaffari, and Breslin 2017) and has been shown to perform well when compared with other metrics (Plank and Van Noord 2011). We would expect a larger JSD value for training and test sets from different content domains than more similar ones. This is borne out by comparing the narrow change JSD values to those of the broad change, as well as comparisons within the narrow change experiments. In particular, all the broad change JSD values are higher than the narrow change JSD values. Moreover, the valence changes from the narrow change DOS yield lower JSD values, while the Ess comparisons tend to have higher JSD values.

The three classification models used for evaluation rely on either *n*-grams alone (*n-grams*), both *n*-grams and PS-scaled linguistically defined specific detail features (*n-grams* + *spec det*), or PS-scaled specific detail features alone (*spec det*). The *n*-grams model serves as a baseline model comparable to prior work using *n*-grams, while the other two models allow us to see the impact

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 6. Within-domain classification results on the DOS, Ess, and DI corpora. Baseline features are *n*-grams that aren't PS-scaled, while specific detail (spec det) features are PS-scaled. Each column reports precision (*P*), recall (*R*), and *F*-score (*F*)

			n-grams		*n*-grams + spec det		spec det	
Train	Test		T	D	T	D	T	D
DOS	Cross-validation	P	0.85	0.88	0.85	0.89	0.66	0.59
		R	0.89	0.85	0.89	0.85	0.48	0.75
		F	0.87	0.86	0.87	0.87	0.55	0.66
Ess	Cross-validation	P	0.71	0.71	0.70	0.70	0.62	0.57
		R	0.71	0.71	0.70	0.70	0.45	0.72
		F	0.71	0.71	0.70	0.70	0.52	0.64
DI	Cross-validation	P	0.56	0.57	0.54	0.55	0.51	0.50
		R	0.60	0.53	0.57	0.52	0.24	0.77
		F	0.58	0.55	0.56	0.53	0.32	0.61

DOS, Deceptive Opinion Spam; Ess, Essays; DI, Deceptive Interview.

of additionally including linguistically defined specific detail features or relying only on those features.

Following prior work investigating classifiers for deception detection (e.g., Bachenko *et al.* 2008; Mihalcea and Strapparava 2009; Ott *et al.* 2011, 2013; Feng *et al.* 2012; Feng and Hirst 2013; Fornaciari and Poesio 2014; Li *et al.* 2014), we report precision, recall, and *F*-score to assess classifier performance. Precision and recall are particularly important for deception detection because deceptive texts often don't surface in the real world at the same rate as truthful ones (i.e., the classes are unbalanced). So, a classifier must make the trade-off between not letting deceptive texts slip through (thereby increasing the false positives) and not impugning truthful texts (thereby increasing the false negatives). Reporting both precision and recall facilitates the interpretation of our findings for future applications that may prioritize different performance goals.

4.1 Within domain

Classifier performance within domain is shown in Table 6. From this, we can make a few observations. First, performance on the DOS corpus is comparable to prior results for *n*-grams (*F*: 0.86–0.87), but drops precipitously when classification is attempted for the Ess (*F*: 0.71) and DI (*F*: 0.55–0.58) corpora. This is also true when using both *n*-gram and specific detail features (DOS *F*: 0.87; Ess *F*: 0.70; DI *F*: 0.53–0.56). In contrast, using specific details alone fairs relatively poorly over all three data sets (DOS *F*: 0.55–0.66; Ess *F*: 0.52–0.64; DI *F*: 0.32–0.61). This suggests that there's no obvious benefit from using the specific detail features, with one notable exception: recall for deceptive documents. Recall over deceptive documents for all three data sets is consistently higher when relying on specific details alone (DOS: 0.75, Ess: 0.72, DI: 0.77). While this is lower than the approaches involving *n*-grams for the DOS corpus (Deceptive R: 0.85; one-sided *t*-test $p = 1\times10^{-6}$), it's equivalent for the Ess corpus (Deceptive R: 0.70–0.71; one-sided *t*-test $p = 0.46$) and markedly higher for the DI corpus (Deceptive R: 0.52–0.53; one-sided *t*-test $p = 0.004$).

Interestingly, this pattern seems correlated with how broad the topics within each corpus are. The DOS is the narrowest, consisting of hotel reviews and the specific details are least beneficial to deceptive recall; the DI is the broadest of all, covering 12 distinct interview questions, and the specific details are most beneficial to deceptive recall. This is a pattern that will reoccur when we examine cross-domain deception detection performance.

Table 7. Narrow change classification results on the DOS, Ess, and DI corpora. Baseline features are *n*-grams that aren't PS-scaled, while specific detail (spec det) features are PS-scaled. Each column reports precision (*P*), recall (*R*), and *F*-score (*F*). Significantly better results versus the *n*-grams baseline are bolded

			n-grams		*n*-grams + spec det		spec det	
Train	Test		T	D	T	D	T	D
DOS+	DOS−	P	0.69	0.84	0.7	0.83	0.62	0.59
		R	0.89	0.60	0.87	0.62	0.54	**0.67**
		F	0.78	0.70	0.77	0.71	0.58	0.63
DOS−	DOS+	P	0.72	0.92	**0.73**	0.92	0.70	0.58
		R	0.95	0.63	0.94	**0.65**	0.39	**0.83**
		F	0.82	0.75	0.82	**0.76**	0.50	0.68
Ess-Ab+BF	Ess-DP	P	0.65	0.59	0.65	0.61	0.58	0.63
		R	0.50	0.73	0.55	0.72	**0.72**	0.48
		F	0.56	0.65	0.6	0.66	0.64	0.54
Ess-Ab+DP	Ess-BF	P	0.60	0.57	0.62	0.58	0.66	0.59
		R	0.51	0.66	0.49	0.70	0.48	0.75
		F	0.55	0.61	0.54	0.63	0.55	0.66
Ess-BF+DP	Ess-Ab	P	0.73	0.64	0.75	0.65	0.60	0.53
		R	0.57	0.79	0.57	0.81	0.29	0.81
		F	0.64	0.71	0.64	0.72	0.39	0.64

DOS, Deceptive Opinion Spam; Ess, Essays; DI, Deceptive Interview; Ess-Ab, Essays corpus about abortion; Ess-BF, Essays corpus about best friends; Ess-DP, Essays corpus about the death penalty.

Still, for best overall performance within domain as measured by *F*-score, using *n*-grams alone seems the best. That is, using specific details is rarely helpful if training data are available from the domain in which we wish to detect deception.

4.2 Narrow change results

Classifier performance with a narrow content change between training and test sets is shown in Table 7. We note that both narrow and broad change classification results use a single training and test set, unlike the within-domain cross-validation setup. Therefore, we use a bootstrap resampling test (Graham, Mathur, and Baldwin 2014) to check significance. Significantly better results for *ngrams + spec det* and *spec det* versus the *n-grams* baseline are bolded in Tables 7 and 8. From this, we can again make a few qualitative observations. First, as mentioned above, the within data set content changes differ: DOS only changes the valence of the hotel reviews, while Ess has a broader topic change. This has a strong impact on the *F*-score performance for the classifiers using *n*-grams alone and *n*-grams with specific details, with the DOS scores markedly higher (DOS F: 0.70–0.82; Ess F: 0.54–0.72).

Relatedly, this difference in content change impacts the utility of incorporating specific detail features. For DOS valence changes, including specific detail features along with *n*-gram features doesn't improve *F*-scores (*n*-grams: 0.70–0.82; *n*-grams + specific details: 0.71–0.82), and relying on specific details alone is notably worse, especially for accurately classifying truthful documents (specific details Truthful: 0.50–0.58; Deceptive: 0.63–0.68). However, recall on deceptive documents is significantly better when incorporating specific details (DOS + training, specific details: 0.67 vs. 0.60–0.62; DOS− training, specific details: 0.83 and *n*-grams + specific details: 0.65 vs.

Table 8. Broad change classification results on the DOS, Ess, and DI corpora. Baseline features are *n*-grams that aren't PS-scaled, while specific detail (spec det) features are PS-scaled. Each column reports precision (*P*), recall (*R*), and *F*-score (*F*). Significantly better results versus the *n*-grams baseline are bolded

			n-grams		*n*-grams + spec det		spec det	
Train	Test		T	D	T	D	T	D
DOS + Ess	DI	P	0.53	0.56	0.54	0.56	0.55	0.51
		R	0.70	0.37	0.67	**0.42**	0.22	**0.81**
		F	0.60	0.44	0.60	**0.48**	0.32	**0.63**
Ess + DI	DOS	P	0.51	0.53	0.51	0.56	**0.54**	0.55
		R	0.72	0.31	**0.85**	0.18	0.61	**0.49**
		F	0.60	0.39	**0.64**	0.28	0.57	**0.52**
DOS + DI	Ess	P	0.51	0.55	0.51	0.56	**0.61**	0.53
		R	0.84	0.19	0.84	0.20	0.27	**0.82**
		F	0.63	0.29	0.63	0.29	0.37	**0.64**

DOS, Deceptive Opinion Spam; Ess, Essays; DI, Deceptive Interview.

n-grams: 0.63). This boost can also yield a higher *F*-score on deceptive documents when using specific details (DOS−, *n*-grams + specific details: 0.76 vs. *n*-grams 0.75).

Moreover, for the Ess topic changes, relying on specific details alone is often about as good as relying on *n*-grams or both *n*-grams and specific details (e.g., Ess-Ab + DP Truthful: *n*-grams = 0.55; *n*-grams + specific details = 0.54; specific details = 0.55). In fact, there are a few cases where relying on specific details alone is better (e.g., Ess-Ab + DP Deceptive: *n*-grams = 0.61, *n*-grams + specific details = 0.63, specific details = 0.66), though also ones where it's markedly worse (e.g., Ess-BF + DP Truthful: *n*-grams = 0.64, *n*-grams + specific details = 0.64, specific details = 0.39). Interestingly, there's also one case where specific details alone are better for recall on truthful documents: Ess-Ab + BF specific details: 0.72 vs. 0.50–0.55 for the other approaches.

As in the within-domain results, we also see a trend where relying on specific details alone often improves the recall on deceptive documents (e.g., DOS+: *n*-grams = 0.63, *n*-grams + specific details = 0.65, specific details = 0.83; and Ess-Ab + DP: *n*-grams = 0.66, *n*-grams + specific details = 0.60, specific details = 0.75). Yet, when this occurs, there is also a noticeable drop in Truthful recall (e.g., DOS+: *n*-grams = 0.95, *n*-grams + specific details = 0.94, specific details = 0.39; and Ess-Ab + DP: *n*-grams = 0.51, *n*-grams + specific details = 0.49, specific details = 0.48).

We interpret these results generally to indicate that specific details tend to have added value the more the training data differ in content from the test data. However, the benefit appears primarily in accurately detecting deceptive documents, at the cost of lower accuracy on truthful documents. We would therefore predict that when the content changes even more broadly from training to test sets, specific details should be even more valuable for detecting deception. Still, the overall detection performance is noticeably lower across domain than within domain, no matter which features are used.

4.3 Broad change results

Classifier performance with a broad content change between training and test sets is shown in Table 8. From these, we again make a few qualitative observations. First, we see here a more extreme *F*-score performance drop, with the highest *F*-score of any classifier at 0.64 (Ess + DI Truthful, *n*-grams + specific details). This underscores the difficulty of cross-domain deception detection when the content differs dramatically.

Second, there are distinct performance patterns when *n*-grams are involved (*n*-grams, *n*-grams + specific details) versus when they aren't (specific details alone). When the classifier features involve *n*-grams, we see much higher Truthful recall than Deceptive recall (Truthful: 0.67–0.85, Deceptive: 0.18–0.42), which leads to higher Truthful *F*-scores than Deceptive *F*-scores (Truthful: 0.60–0.64; Deceptive: 0.28–0.48). This contrasts with classifiers relying on specific detail features alone, which either have recall scores that are closer between Truthful and Deceptive Documents (Ess + DI, Truthful: 0.61, Deceptive: 0.49) or have much higher Deceptive recall scores (DOS + Ess and DOS+DI Truthful: 0.22–0.27, Deceptive: 0.81–0.82). This leads to either more balanced *F*-scores (Ess + DI Truthful: 0.57, Deceptive: 0.52) or much higher Deceptive performance (DOS + Ess and DOS + DI Truthful: 0.32–0.37, Deceptive: 0.63–0.64). Overall, relying on specific details shows significant gains primarily in deceptive document recall (as indicated by the bolded values in Table 8), though occasionally truthful document recall is also significantly improved.

We interpret these results to mean that when there are broad content differences across the training and test set domains, relying on specific detail features alone may have an advantage. This advantage primarily surfaces as increased performance on accurately classifying deceptive documents. However, when the deceptive performance increases the most, the truthful performance seems to decrease the most. So, if false negatives (missing deceptive documents) are more costly than false positives (incorrectly identifying something as deceptive when it's not), then specific details alone may be the better choice.

4.4 Utility of linguistically defined specific detail features

4.4.1 Feature weights for n-grams+specific details

We can also evaluate the utility of our linguistically defined specific detail features by seeing if they were used by the classifier when it had a choice between *n*-gram features and the specific detail features. In particular, the statistical *t*-test analysis in Table 4 identified that certain linguistically defined specific detail features differed more strongly when it came to mean values between truthful and deceptive documents across all three data sets. That is, these features differed in general in their values across collections of truthful and deceptive documents. However, as mentioned before, a classifier is capable of detecting more subtle patterns than this and must classify each individual document on the basis of its particular feature values. Therefore, given the statistical promise of these features for classifying an individual document accurately, did a classifier that could rely on either *n*-grams or these specific detail features (the *n-grams+spec det* option) actually use any of these specific detail features? If so, this is additional support for these features' utility for a classification task; if not, this suggests *n*-grams may be preferable to them. Table 9 lists specific detail features appearing in the top 1% of classifier features by weight when detecting deception across domains that differed either narrowly or broadly. We interpret any feature this heavily weighted to be of significant use to the classifier.

We can make a few observations. First, the Numbers feature (which had significantly different mean values between truthful and deceptive texts) is typically used by the classifier: it appears as a top feature for six out of the eight domain change scenarios in Table 9 (column F5). Moreover, for the three broad change scenarios, it's always in the top features. This suggests Numbers is a generally relied-upon feature when detecting deception across domains that differ more starkly in content.

Similarly, features based on prepositions (PP count and PP length, columns F1 and F2 in Table 9) appear in the top features fairly often. For instance, in all the narrow change scenarios, one or the other PP-based feature is in the top features. This aligns with the statistical analysis based on mean values, which found that there are statistically significant differences between truthful and deceptive texts.

However, features involving adjectives (AdjP count and AdjP length, columns F3 and F4 in Table 9) don't appear to be as useful. They're never in the top features for the classifier. This is

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table 9. Presence of our specific detail features in the top 1% of features for each narrow and broad change classification, as determined by the SVM classifier when using both *n*-grams and specific detail features. Specific detail features are indicated by F#: F1 = PP count, F2 = PP length, F3 = AdjP count, F4 = AdjP length, F5 = Numbers, F6 = Proper Nouns, and F7 = Consecutive Nouns. Relative ranking for features within the top 1% is indicated

	Train	Test	F1	F2	F3	F4	F5	F6	F7
Narrow change	DOS+	DOS−	3				2		1
	DOS−	DOS+		2			1		
	Ess-Ab + BF	Ess-DP		1					
	Ess-Ab + DP	Ess-BF	1					2	
	Ess-BF + DP	Ess-Ab	1				2		
Broad change	DOS + Ess	DI	3	2			1		
	Ess + DI	DOS					2		1
	DOS + DI	Ess					1		

SVM, support vector machine, PP, prepositional phrase; AdjP, adjective phrase; DOS+, Deceptive Opinion Spam positive; DOS−, Deceptive Opinion Spam negative; Ess-Ab, Essays corpus about abortion; Ess-BF, Essays corpus about best friends; Ess-DP, Essays corpus about the death penalty; DI, Deceptive Interview.

somewhat unexpected, given that (i) they have statistically different mean values between truthful and deceptive texts, and (ii) adjective phrases intuitively capture a way to add more detail. Regarding this second point about how adjectives may be an intuitive way to add more detail, perhaps features involving adjectives are also easier to manipulate consciously when fabricating text. That is, perhaps deceivers more often draw on adjective phrases when constructing deceptive reviews precisely because adjectives are an intuitive way to add more detail. This would then cause adjective phrases to differ less between truthful and deceptive texts.

Regarding the statistical difference in mean values, it may be that the statistically different mean values emerge for adjectives only when content from all three domains is considered together. That is, looking at mean values of truthful versus deceptive texts in opinion spam, essays, and interview answers collectively is different than training on truthful and deceptive in some domains (e.g., opinion spam and essays) and testing on truthful and deceptive in a different domain (e.g., interview answers). So, it's possible that if we looked at mean values of adjective-based features in the specific narrow change and broad change evaluation scenarios we investigated, the mean values would no longer differ so much.

This seems to be somewhat true when we analyze the training sets of the narrow and broad change scenarios, which is what the classifier uses to set weights for the different available features. In particular, only the training sets that involve the DOS corpus (i.e., DOS+, DOS−, or the entire DOS corpus) show mean values of adjective-based features that are significantly different between truthful and deceptive texts. For example, while the narrow change scenarios with DOS have low p values based on two-sample, independent t tests (e.g., DOS training, AdjP Count: $p = 0.002$–0.047; AdjP Length: $p = 0.011$–0.034), the narrow change scenarios with the Ess have much higher p values (Ess training, AdjP Count: $p = 0.17$–0.46; AdjP Length: $p = 0.22$–0.89). This suggests that adjective-based features wouldn't obviously be diagnostic for scenarios that don't involve the DOS corpus in the training set.

For the scenarios that do involve DOS data in the training set, we might expect that features rank in the top 1% if their mean values are highly distinctive between truthful and deceptive texts. An analysis of the broad change scenarios suggests that this is mostly true. For example, when we consider how adjective-based features compare to the other feature included in the classifier's top 1% for the DOS + DI broad change scenario (Numbers), we find that Numbers have mean values that are far more distinctive than those of the adjective-based features (Numbers: two-sample, independent t-test $p = 1.8\times10^{-20}$; AdjP count: $p = 0.00017$; AdjP length: $p = 0.00081$). This suggests that while adjective-based features are potentially diagnostic, they aren't *as* diagnostic as the Numbers feature.

When we consider the other broad change scenario that involves the DOS corpus in its training set (DOS + Ess), we find similar—though not identical—results. In particular, there are three features in the classifier's top 1%: Numbers, PP length, and PP count. These features do have mean values whose two-sample independent t-test p values suggest that the feature mean values are quite distinctive between truthful and deceptive texts (Numbers: $p = 7.0 \times 10^{-27}$, PP length: $p = 0.00035$; PP count: $p = 0.0031$). However, if mean value distinctiveness were the only factor in the classifier's weighting, we would expect any feature whose means had a p value less than 0.0031 to also be included in the top 1%. It turns out that *both* adjective-based features have p values less than this threshold for this training set (AdjP count: $p = 0.00012$, AdjP length: $p = 0.0011$). So, it may be that the classifier is relying on more subtle patterns than simple mean value of a feature when making its classification decisions.

Interestingly, the two noun-based features, Proper Nouns (Column F6) and Consecutive Nouns (Column F7), do appear in the top features occasionally. Notably, when consecutive nouns are in the top features, they are *the* top-ranked specific detail feature (DOS+, Ess+DI). This may be somewhat surprising, given how rarely these features had statistically significant mean values between truthful and deceptive texts (see Table 4). However, it seems that they sometimes have other quantitative patterns the classifier can leverage when identifying if a particular text is truthful or deceptive. Moreover, it may well be that proper nouns would occur more often in the kinds of real-world deception detection scenarios that cross-domain work will be aimed at, such as court testimony, police interrogations, and depositions.[h] So, the utility of proper nouns for distinguishing truthful from deceptive texts is a welcome finding.

More generally, we interpret these findings to indicate that certain linguistically defined specific detail features are useful to classifiers detecting deception across domain. In particular, the specific detail features based on numbers are most relied upon, with preposition-based features sometimes relied upon, and noun-based features occasionally relied upon.

4.4.2 Feature weights for specific details alone

The most striking performance difference for the broad content change scenarios in Table 8 was that a classifier relying on specific details alone (the *spec det* option) had markedly better performance on deceptive texts, due to its much higher recall on these texts. Another test of specific detail feature utility is to see the relative weighting of the features that led to this distinct behavior pattern.

For all three broad change scenarios, the Numbers feature is most relied upon, with a weight that is typically an order of magnitude higher than the weights of the other specific detail features. This suggests the Numbers feature is important for generating this deception detection performance pattern. Additionally, the four top-weighted features for all three scenarios always include the Consecutive Nouns feature, and one (Ess + DI) also includes the Proper Nouns feature. This suggests that noun-based features are also more relied upon to generate this notable performance pattern. Two other features also occur in the top-weighted four features of at least two broad change scenarios: PP Length and AdjP Count. Recall that while preposition-based features were relied upon by a classifier that could use n-grams or specific details, neither adjective-based feature was. Here, we see that that length of PPs and the number of adjective phrases both seem to contribute to the higher performance on deceptive texts. Therefore, adjective-based features may be useful precisely when accurate performance on deceptive texts is prioritized.

[h] We thank an anonymous reviewer for noting this.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

5. Discussion

5.1 Overview

We have investigated linguistically defined specific details for cross-domain deception detection, defining these features by synthesizing insights from previous approaches. More specifically, we derived our features by considering (i) that linguistically defined features are more likely to be generalizable across different content areas, and (ii) specific details can be a linguistic reflection of the psychological process underlying truthful versus deceptive content generation.

We then evaluated the utility of these features on corpora that allowed us to vary how much the content changes between training and test sets. The changes ranged from fairly minimal (only the valence changed) to quite broad (from noninteractive reviews of hotels and emotionally charged opinions to interactive job interview questions).

Our key finding is that linguistically defined specific detail features are most useful when the content changes are more dramatic between training and test. This accords with the intuition that linguistically defined features are more generalizable across domains. However, we note that there is a significant drop in overall classification performance the more the content varies between training and test—the best F-score for any classifier is 0.91 when the content doesn't change at all, but the best F-score is 0.64 when the content changes most broadly. Additionally, the benefit of linguistically defined specific detail features is for classifying deceptive texts: there are fewer false negatives than when n-gram features are used for classification. However, this comes at the cost of poorer classification on truthful texts. So, if false negatives (e.g., letting false texts slip through) are more costly than false positives (e.g., incorrectly tagging true texts as false), then these linguistically defined specific detail features are beneficial.

Interestingly, the particular specific detail features we defined vary with respect to how much a classifier relies upon them. That is, not all features we investigated appear to be equally useful. A notable standout is the Numbers feature (corresponding to uses of exact number words like *two*), which was often in the top-weighted features and often far more heavily weighted than the other features. This feature has the additional benefit of being particularly easy to identify and extract, and so may be an especially good feature to incorporate into future cross-domain deception detection approaches. Future theoretically motivated work can investigate what it is about exact numbers that makes them the most useful specific detail type for detecting deceptive opinions. If specific conceptual or psychological features about exact numbers can be identified, then these features may spur ideas for other linguistic instantiations of specific details not considered here.

Another interesting avenue for future research concerns incorporating more abstract representations of meaning. More specifically, we looked here at generalizable cues based on linguistic structure—instead of looking at individual words, we harnessed syntactic structures of different kinds. Another option is to consider distributed meaning representations, such as word2vec (Mikolov *et al.* 2013), GLoVe (Pennington, Socher, and Manning 2014), and ELMo (Peters *et al.* 2018), that allow more generalizable cues based on linguistic meaning; these representations allow this because vectorized meaning representations can encode meaning in a way that enables useful semantic generalizations across individual words (e.g., recognizing that *awesome* and *fantastic* have similar meanings). The n-gram approaches we investigated here are based on individual words, and so miss out on this kind of semantic relatedness—for example, *is+so+awesome* would be counted as a completely separate n-gram from *are+really+fantastic*, even though they're similar semantically. Approaches relying on distributed meaning representations could therefore offer n-gram-based features that have more semantic generalization, and so might perform better across domains when individual words change. This would be in line with prior work (Pearl and Enverga 2014) that finds improved text classification performance by using more semantically "deep" n-gram features.

Future work on other available cross-domain data sets can also investigate if the results we found here regarding the utility of linguistically defined specific detail features hold up when the content varies even more widely (Pérez-Rosas and Mihalcea 2015) and when the content is generated in real-life scenarios (Bachenko *et al.* 2008; Fornaciari and Poesio 2011, 2013; Burgoon *et al.* 2016).

5.2 Recommendations for cross-domain deception detection

More broadly, our findings suggest specific recommendations for future cross-domain approaches. First, we should expect a nontrivial drop in classification performance when going from within-domain to cross-domain detection, no matter what approach is used. Relying on the specific details we investigated here can't offset that performance drop. So, if it's possible to use training data that's close in domain to the testing data, that's worthwhile (and the closer, the better).

However, of course sometimes this isn't possible. In these cases—and especially when the training content differs significantly from the test content—it's important to weigh the cost of false negatives versus false positives. If false negatives are worse, it's useful to rely on linguistically defined specific details alone; if instead false positives are worse, it's better to rely on a mix of *n*-gram and linguistically defined specific detail features.

6. Conclusion

By synthesizing insights from previous approaches to deception detection, both within-domain and cross-domain, we have linguistically defined specific detail features that are useful at detecting deception across domains in certain scenarios. In particular, our linguistically defined specific details are most helpful when there are dramatic differences in content between training and test sets, and when false negatives are more costly than false positives. We hope future work can harness these insights to improve general-purpose approaches to deception detection.

Acknowledgement: We are especially grateful for the comments and support of the Computation of Language Laboratory and the Summer Undergraduate Research Opportunities Program at UC Irvine. We have additionally benefited from the comments of several anonymous reviewers.

Financial support. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Almela, Á., Valencia-García, R. and Cantos, P. (2012). Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the ACL Workshop on Computational Approaches to Deception Detection*, pp. 15–22.

Bachenko, J., Fitzpatrick, E. and Schonwetter, M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)-Volume 1*, Stroudsburg, PA: Association for Computational Linguistics. pp. 41–48.

Burgoon, J., Mayew, W.J., Giboney, J.S., Elkins, A.C., Moffitt, K., Dorn, B., Byrd, M. and Spitzley, L. (2016). Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. *Journal of Language and Social Psychology* **35**(2), 123–157.

Burgoon, J.K., Blair, J.P., Qin, T. and Nunamaker, J.F. (2003). Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pp. 91–101.

Burgoon, J.K., Buller, D.B., Guerrero, L.K., Afifi, W.A. and Feldman, C.M. (1996). Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communications Monographs* **63**(1), 50–69.

Burgoon, J.K., Buller, D.B., White, C.H., Afifi, W. and Buslig, A.L.S. (1999). The role of conversational involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin* **25**(6), 669–686.

Burgoon, J.K. and Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology* **25**(1), 76–96.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Feng, S., Banerjee, R. and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 171–175.

Feng, V.W. and Hirst, G. (2013). Detecting deceptive opinions with profile compatibility. In *International Joint Conference on Natural Language Processing*, pp. 338–346.

Fitzpatrick, E. and Bachenko, J. (2009). Building a forensic corpus to test language-based indicators of deception. *Language and Computers* **71**(1), 183–196.

Fitzpatrick, E., Bachenko, J. and Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies* **8**(3), 1–119.

Fornaciari, T. and Poesio, M. (2011). Lexical vs. surface features in deceptive language analysis. In *Proceedings of the ICAIL 2011 Workshop: Applying Human Language Technology to the Law*, pp. 2–8.

Fornaciari, T. and Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law* **21**(3), 303–340.

Fornaciari, T. and Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the Association for Computational Linguistics*, pp. 279–287.

Fusilier, D.H., Montes-y-Gómez, M., Rosso, P. and Cabrera, R.G. (2015). Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management* **51**(4), 433–443.

Graham, Y., Mathur, N. and Baldwin, T. (2014). Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 266–274.

Hirschberg, J., Benus, S., Brenier, J., Enos, F., Hoffman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, L.B, Shriberg, E. and Stolcke, A. (2005). Distinguishing deceptive from non-deceptive speech. In *9th European Conference on Speech Communication and Technology*, pp. 1833–1836.

Johnson, M.K. and Raye, C.L. (1981). Reality monitoring. *Psychological Review* **88**(1), 67.

Kim, S., Lee, S., Park, D. and Kang, J. (2017). Constructing and evaluating a novel crowdsourcing-based paraphrased opinion spam dataset. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 827–836.

Klein, D. and Manning, C.D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 423–430.

Kleinberg, B., Mozes, M., Arntz, A. and Verschuere, B. (2017). Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences* **63**(3), 714–723.

Krüger, K., Lukowiak, A., Sonntag, J., Warzecha, S. & Stede, M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering* **23**(5), 687–707.

Larcker, D.F. and Zakolyukina, A.A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research* **50**(2), 495–540.

Levine, T.R. (2014). Truth-Default Theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology* **33**(4), 378–392.

Li, J., Ott, M., Cardie, C. and Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1566–1576.

Mayzlin, D., Dover, Y. and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104**(8), 2421–2455.

McCornack, S.A. (1992). Information manipulation theory. *Communications Monographs* **59**(1), 1–16.

McCornack, S.A. and Parks, M.R. (1986). Deception detection and relationship development: The other side of trust. *Annals of the International Communication Association* **9**(1), 377–389.

Mihalcea, R. and Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 309–312.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119. https://www.cambridge.org/core/journals/natural-language-engineering/article/word2vec/B84AE4446BD47F48847B4904F0B36E0B.

Narayan, R., Rout, J.K. and Jena, S.K. (2018). Review spam detection using opinion mining. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pp. 273–279.

Newman, M.L., Pennebaker, J.W., Berry, D.S. and Richards, J.M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* **29**(5), 665–675.

Ott, M., Cardie, C. and Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 201–210.

Ott, M., Cardie, C. and Hancock, J.T. (2013). Negative deceptive opinion spam. In *HLT-NAACL*, pp. 497–501.

Ott, M., Choi, Y., Cardie, C. and Hancock, J.T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 309–319.

Pearl, L., Lu, K. and Haghighi, A. (2016). The character in the letter: Epistolary attribution in Samuel Richardsons Clarissa. *Digital Scholarship in the Humanities* **32**(2), 355–376.

Pearl, L. and Steyvers, M. (2012). Detecting authorship deception: A supervised machine learning approach using author writeprints. *Literary and Linguistic Computing* **27**(2), 183–196.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Pearl, L.S. and Enverga, I. (2014). Can you read my mindprint?: Automatically identifying mental states from language text using deeper linguistic features. *Interaction Studies* **15**(3), 359–387.

Pennebaker, J., Booth, R. and Francis, M. (2007). *Linguistic Inquiry and Word Count: LIWC*. Austin, TX: LIWC.net.

Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pérez-Rosas, V. and Mihalcea, R. (2015). Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1120–1125.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Plank, B. and Van Noord, G. (2011). Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1566–1576.

Remus, R. (2012). Domain adaptation using domain similarity and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW)*, pp. 717–723.

Rosso, P. and Cagnina, L.C. (2017). Deception detection and opinion spam. In Cambria E., Das D., Bandyopadhyay S. and Feraco A. (eds.), *A Practical Guide to Sentiment Analysis*. Socio-Affective Computing, vol 5. Cham: Springer, p. 155–171.

Rubin, V.L. and Vashchilko, T. (2012). Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pp. 97–106.

Ruder, S., Ghaffari, P. and Breslin, J.G. (2017). Data selection strategies for multi-domain sentiment analysis. arXiv preprint arXiv:1702.02426.

Santos, E. and Li, D. (2010). On deception detection in multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **40**(2), 224–235.

Steller, M. and Koehnken, G. (1989). Criteria-based statement analysis.

Vrij, A. (2000). *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. New York: Wiley.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. New York: Wiley.

Yancheva, M. and Rudzicz, F. (2013). Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 944–953.

Yoo, K.-H. and Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. In: Höpken W., Gretzel U. and Law R. (eds.) *Information and Communication Technologies in Tourism*, Vienna: Springer, pp. 37–47.

Yu, D., Tyshchuk, Y., Ji, H. and Wallace, W. (2015). Detecting deceptive groups using conversations and network analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, pp. 857–866.

Zhou, L., Burgoon, J.K., Nunamaker, J.F. and Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation* **13**(1), 81–106.

Zuckerman, M., DePaulo, B.M. and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology* **14**, 1–59.

Appendix A. Examples from corpora

Table A1. Sample truthful and deceptive positive online reviews from the Deceptive Opinion Spam corpus

Truthful	Deceptive
I only stayed out with my boyfriend for one night, however enjoyed my stay. The staff was friendly, the room was nice and clean, the hallways and ballrooms etc were elegant. Room service was quick and had good options to choose from that actually tasted great. The staff was able to extend our check out time for an extra 1–2 hours without an extra charge to the room. Great location too! Walking distance from the Art Museum, Millennium Park, Grant Park (right across the street) and a quick cab ride to McCormick Place. If I were in the city again I would love to stay there again.	The Hilton in Chicago was awesome. The room was very clean and the hotel staff was very professional. One of the features I liked, was that in my room the internet access was wire and wireless, considering my laptop is not wireless, it help me out alot. Food was very good, quality was great. There was also a flat screen in my room...awesome. The hotel itself is locaated in the middle of alot of resturants with fin dinning. I also enjoyed the gym very much. Overall, I enjoyed myself, and I will stay again at the Hilton when I return to Chicago.

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press

Table A2. Sample statements from true and deceptive essays in the Essays corpus

TRUTHFUL	DECEPTIVE
ABORTION	
I believe abortion is not an option. Once a life has been conceived, it is precious. No one has the right to decide to end it. Life begins at conception, because without conception, there is no life.	A woman has free will and free choice over what goes on in her body. If the child has not been born, it is under her control. Often the circumstances an unwanted child is born into are worse than death. The mother has the responsibility to choose the best course for her child.
DEATH PENALTY	
I stand against death penalty. It is pompous of anyone to think that they have the right to take life. No court of law can eliminate all possibilities of doubt. Also, some circumstances may have pushed a person to commit a crime that would otherwise merit severe punishment.	Death penalty is very important as a deterrent against crime. We live in a society, not as individuals. This imposes some restrictions on our actions. If a person doesn't adhere to these restrictions, he or she forfeits her life. Why should taxpayers' money be spent on feeding murderers?
BEST FRIEND	
I have been best friends with Jessica for about seven years now. She has always been there to help me out. She was even in the delivery room with me when I had my daughter. She was also one of the Bridesmaids in my wedding. She lives six hours away, but if we need each other we'll make the drive without even thinking.	I have been friends with Pam for almost four years now. She's the sweetest person I know. Whenever we need help she's always there to lend a hand. She always has a kind word to say and has a warm heart. She is my inspiration.

Table A3. Questions from the Deceptive Interview corpus

#	Question
1	Please describe your educational background.
2	If you found a wallet containing $1000 and no identification, what would you do with it? Why?
3	What do you consider to be your biggest strengths?
4	Please describe your current or last occupation.
5	What types of people tend to rub you the "wrong way"?
6	You told me that your most recent occupation was as a ________. Please describe a typical day for you as a ________.
7	What political issues do you feel strongly about and why?
8	If your best friend was cheating on his or her spouse, what would you do? Why?
9	With which political party do you most closely identify and how does that fit with your political views?
10	How ambitious are you?
11	Describe the last argument or fight you had with a close friend or family member.
12	How happy are you with your life right now and why?

Table A4. Sample true and deceptive responses to *"Please describe your educational background."* from the Deceptive Interview corpus

Truthful	Deceptive
I have a bachelors of arts in education. I have an associates degree in accounting and computerized, eh um, bookkeeping and I have an artisans training in crafts. About eighteen years of formal school and about 45 years of practice. Oh yes, very much so. Um, not necessarily, I think a person who wants to be a teacher has to be very much dedicated, now more than ever. And as for accounting, that is just wisdom in these economic times. And I happen to be a creative fidget when it comes to crafts.	Well, I am a, I completed my masters degree in business administration. And I am hopefully going to be completing one for my doctorate, depending on time and money. In December of 1990. U of A. As I say that depends on money and the family situation. When I have time and money and work allows and everything else. Where did I complete that, I did that in '87, and I took some time off and went back. Here in Tucson.

Appendix B. Identifying specific details manually

Table A5. A positive truthful review and a positive deceptive review of the James Hotel from the Deceptive Opinion Spam corpus. Linguistic expressions that seemed upon manual inspection to encode specific details are bolded. These kind of manually identified specific detail examples were used to inspire the linguistic definitions of specific details

Truthful
The James Hotel met and exceeded our expectations. A haven of **cool, uncluttered comfort in a hot, crowded but congenial city**. **Located within steps of some of the greatest art, architecture, culinary and cultural opportunities anywhere in North America**, the staff of the James provided us with a home away from home for the week we visited. **The check-in was quick and flawless** and **the king suite a calm, restorative oasis**. It is difficult to over-praise the friendly, attentive staff; this is a hotel that truly has its act together! **The day before our visit ended, we experienced problems with the room's air conditioner**. When the technician was unable to quickly solve the problem, we were immediately upgraded to a **one-bedroom apartment with stunning views of the Chicago skyline**. We eagerly await the opening of new James Hotels in other cities.
Deceptive
The James Chicago Hotel is located right in the heart of the one and only, downtown Chicago. This wonderful hotel has **many classy and down to earth boutique's**, and the views are breath-taking. The hotel just has a warm feeling to it, and the hotel staff is more than excellent. This hotel is situated **right across from the Magnificent Mile**, and only **minutes away from the world famous shopping, dining, and fun for all ages**. The James is modern and luxurious. The James is definitely a hotel that I would recommend and will go back to in years to come.

Cite this article: Vogler N and Pearl L (2020). Using linguistically defined specific details to detect deception across domains. *Natural Language Engineering* **26**, 349–373. https://doi.org/10.1017/S1351324919000408

https://doi.org/10.1017/S1351324919000408 Published online by Cambridge University Press