Quotation Detection and Classification with a Corpus-Agnostic Model

Sean Papay & Sebastian Padó

Institute for Natural Language Processing
University of Stuttgart, Germany
{sean.papay,pado}@ims.uni-stuttgart.de

Abstract

The detection of quotations (i.e., reported speech, thought, and writing) has established itself as an NLP analysis task. However, state-of-the-art models have been developed on the basis of specific corpora and incorporate a high degree of corpus-specific assumptions and knowledge, which leads to fragmentation. In the spirit of task-agnostic modeling, we present a corpus-agnostic neural model for quotation detection and evaluate it on three corpora that vary in language, text genre, and structural assumptions. The model (a) approaches the state-of-the-art on the corpora when using established feature sets and (b) shows reasonable performance even when using solely word forms, which makes it applicable for non-standard (i.e., historical) corpora.

1 Introduction

Quotation is a general notion that covers different kinds of direct and indirect speech, thought, and writing in text (Semino and Short, 2004). Quotations are a prominent linguistic device used to express claims, assessments, or attitudes attributed to speakers. Consequently, the analysis of quotations is gaining traction in computational linguistics and digital humanities, providing evidence for speaker relationships (Elson et al., 2010; Agarwal et al., 2012), inter-speaker sentiment (Nalisnick and Baird, 2013), politeness (Faruqui and Pado, 2012), and narrative structure (Jannidis et al., 2018).

As is often the case with semantic phenomena, manual annotation of quotations has shown to be slow and resource-intensive, in particular when undertaken in conjunction with the annotation of speakers and information quality (Brunner, 2013; Pareti, 2015). This provides the rationale for automatic *quotation recognition* methods. After a first round of rule-based methods (Pouliquen et al., 2007; Brunner, 2013), recent supervised models use mostly sequence classifiers (Pareti et al., 2013; Almeida et al., 2014; Scheible et al., 2016).

Not surprisingly, these corpora differ substantially across a number of relevant dimensions. including text genre, annotation scheme, and theoretical assumptions. For example, Pareti et al. (2013) focus exclusively on newspaper text and focus on developing a *uniform* annotation schema that captures the shared properties of all kinds of annotations. Thus, even though this corpus contains direct, indirect, and mixed quotations, these not marked as instances of their specific subtypes. In addition, each quote is assumed to be introduced by a *cue* (markables are shown surrounded by red square brackets):

(1) Hillary Clinton on Saturday [acknowledged]$_{\text{CUE}}$ [the state of the economy is good]$_{\text{QUOTE}}$.

This assumption is generally true for newspaper text, and simplifies the task of quotation detection.

The situation is rather different in the literary texts considered by Semino and Short (2004). Cues are much more varied, and are sometimes omitted entirely, such as in this exchange from Dickens' *Christmas Carol*:

(2) [”Much!”]$_{\text{QUOTE}}$ – Marley's voice, no doubt about it.
[”Who are you?”]$_{\text{QUOTE}}$
[”Ask me who I was.”]$_{\text{QUOTE}}$

The study follows a generally more *differentiating* approach. It develop and annotate a rich typology of different subtypes of quotations to distinguish, e.g., direct from indirect quotations, and speech from thoughts from writing.

Not surprisingly, therefore, all models for quotation detection were developed for one specific corpus. This leads to two problems:

Proceedings of Recent Advances in Natural Language Processing, pages 888–894,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_103

1. The models inherit the corpora's *structural and theoretical assumptions*, such as the presence of a cue assumed by models for the Pareti et al. (2013) corpus.

2. The models typically include *domain-specific* features and knowledge sources that happened to be available from the corpus, such as lists of likely cue verbs or syntactic realizations of quotations.

This corpus dependence amounts to conceptual overfitting: while it leads to better fit for the original corpus, models are not transferable to new domains and analysis schemes. In other words, it leads to serious *fragmentation*.

In this paper, we develop and evaluate a *corpus-agnostic* neural model architecture for automatic quotation recognition that makes as few assumptions as possible about the corpus to be modelled but is still expressive enough to deal with the challenge of recognizing quotation spans of essentially arbitrary length (Scheible et al., 2016). In this respect, we see our study as a step towards transfer learning (Pan and Yang, 2009) and task-agnostic learning (Hashimoto et al., 2017). We find that our model can perform reasonably well across corpora differing in genre, language, and structure.

2 Related Work: Datasets and Models

We now review the state of the art in automatic quotation annotation, describing the three major quotation corpora for English and German and the corresponding models. We exclude corpora that focus on one specific quotation subtype such as the Columbia Speech Attribution corpus (Elson and McKeown, 2010) which only covers direct speech.

2.1 PARC Dataset

Dataset. The Penn Attribution Relation Corpus (Pareti, 2015), version 3 (PARC3) is a subset of the Penn Treebank, annotated with quotations and attribution relations. It consists of English newswire text from the Wall Street Journal. Each attribution relation consists of a cue, optionally a source (speaker), and content (quotation span), all marked as text spans. As part of the Penn Treebank, PARC3 provides manually annotated tokenization, POS tags, lemmas, and constituency parses.

Quotation spans are not labeled with more specific types, but PARC3 distinguishes informally (based on the surface form) between direct quotations (starting and ending with quotation marks), indirect quotations (without any quotation marks), and mixed quotations (everything else).

Pareti model. Pareti (2015), an extension of Pareti et al. (2013), presents a pipeline architecture for quotation annotation. It first applies a k-NN classifier to identify quotation cues within the corpus. Then, a linear-chain conditional random field (CRF) is used to identify quotation spans in the vicinity of each cue. The Pareti model builds on corpus-specific knowledge, including lists of cue verbs, and handcrafted features sensitive to punctuation conventions in English newswire text.

Scheible model. Scheible et al. (2016) retain the pipeline architecture of Pareti (2015) and its feature set, but replace the components. Cue annotation is performed with an averaged perceptron. More importantly, they replace quotation annotation proper with a sampling-based procedure: a perceptron samples tokens as likely span boundaries, which are then combined into complete quotation spans, using a semi-Markov model.

2.2 STOP Dataset

Semino and Short (2004) presents a corpus-based ontology of quotations in English text. It introduces two dimensions: (a), speech vs. thought vs. writing; and (b), direct vs. indirect vs. free indirect vs. reported, yielding a Cartesian product of twelve quotation subclasses. These are used to annotate the Speech, Thought, and Writing Presentation corpus (STOP). It comprises 120 sections, split evenly across three genres (fiction, newspaper, and biographies), of about 2,000 words each (Total size: 250,000 tokens; 8,000 quotations). The corpus has no linguistic annotation: the only features available are words' surface forms. To our knowledge, there are no models for this dataset.

2.3 *Redewiedergabe* Dataset

Dataset The Redewiedergabe ('reported speech') corpus (RWG) (Brunner, 2013) is a corpus of German narrative text, comprising thirteen public-domain German narratives (1787–1913). The quotation annotations in RWG adopt the scheme by Semino and Short (2004) and distinguish direct, indirect, free indirect, and reported variants of speech, thought, and writing. The total size of the corpus is 57.000 tokens, and 17.000 quotation spans.

Unlike STOP, RWG contains some linguistic information, namely POS tags, lemmas, and mor-

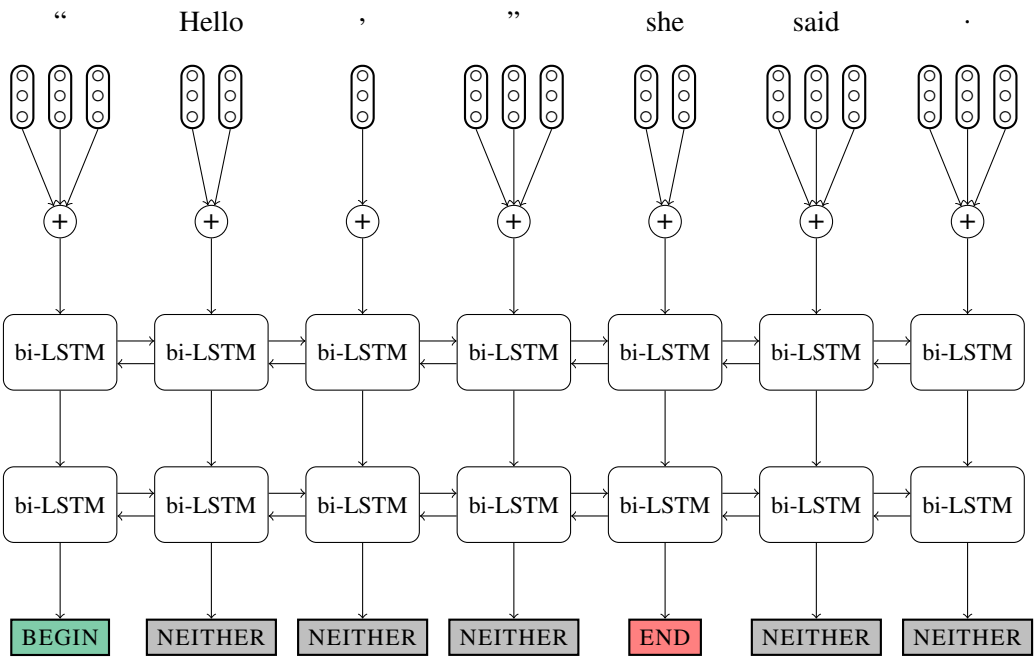

Figure 1: The NQD architecture. Tokens are represented as a bag of feature embeddings, and each token sequence is processed by a 2-layer bi-LSTM network, before a max-entropy classifier labels each token.

phological features (case, number, gender). This information is obtained automatically, though.

Models. Brunner (2013) proposes two models for quotation annotation on RWG. Both models work at the sentence level and predict only the presence of absence of quotations, not their spans (even though this information is annotated). The first model is rule-based (**Brunner RB**). It uses a set of handcrafted rules to identify direct, indirect, reported, and free indirect quotations. The second model (**Brunner ML**) is a simple classification model based on random forests.

3 Neural Quotation Detection (NQD)

We now define a neural architecture, NQD, with the goal of modeling the quotations in all three corpora described in Section 2. We design our model to leverage the commonalities across datasets, while not depending on the features of any dataset in particular. As all datasets involve long quotation spans with long-distance dependencies, an LSTM-based approach was natural, given such models' ability to capture very long-distance dependencies of up to 200 tokens (Khandelwal et al., 2018). Conversely, given the structural differences between corpora, we decided against a pipeline model like those employed by Pareti (2015) and Scheible et al. (2016) which predict cues first and then quotation spans. NQD predicts quotations directly without explicitly identifying cues.

NQD frames quotation prediction as token classification, classifying each token as either beginning a quotation (BEGIN), ending a quotation (END), or neither (NEITHER). Quotation spans then consist of all tokens starting with a BEGIN tag, up to (but not including) the next END or BEGIN tag, or the end of sequence. This model is not limited to the sentence level: it is able to make predictions for a whole document and in this manner can capture very long quotation spans (Scheible et al., 2016). Concretely, the sequence-to-sequence architecture comprises a 2-layer bi-LSTM network, with the outputs of the second bi-LSTM feeding into a 3-class softmax classifier. Thus, the model takes token sequences as input and produces a sequence of token labels. Figure 1 shows a schematic diagram of the NQD architecture. For datasets with multiple quotation types, NQD uses a separate sequence-to-sequence model for each span type, connecting them by weight sharing. All NQD code is available from `https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/index.en.html`.

In the input token sequences, each token is a bag of features. Each feature value is represented as an n-dimensional continuous vector, and each token is represented as the sum of these vectors.

This approach to feature representation allows our model to work with corpora with arbitrary types of token-level features. In the simplest case, when only raw text is present in the corpus, each token is given a single feature for that token's word. If other token-level features are present, such as POS-tags, lemmas, or even parse tree information, these can be incorporated as additional feature vectors, without requiring any changes to the model architecture. Feature vectors can also be initialized to pre-trained representations (e.g. word embeddings) when these are available, or initialized randomly and learned when they are not. Section 2 describes in detail which features are used for the corpora we experiment with.

4 Experimental Evaluation

We now train and test NQD on the three corpora and compare against the state-of-the-art.

4.1 Experimental Setup

PARC3. For PARC3, we train a single classifier on the quote content spans and ignore the cue and source spans. As features, we use token surface forms, lemmas, POS tags, as well as, for each token, the bags of constituents that start with, end with, and contain it. These features are a subset of the features used by Scheible et al. (2016) and Pareti (2015), and like these studies, we use gold standard annotation. We initialize the features for word surface forms with the default GloVe Wikipedia word embeddings (Pennington et al., 2014). Our model makes predictions on entire documents at a time. We use performance on the corpus's development set to guide early stopping during training, and we evaluate on the corpus's test set.

STOP. For STOP, we train four classifiers for the four quote types (direct, indirect, free indirect, reported). We train and evaluate our model on a per-document basis as for PARC3. We use word surface forms (and their GloVe embeddings) as features, we used no other features in this model. As the corpus contains no held-out development or test sets, we used 10-fold cross validation to evaluate our model, using 8 folds for training, 1 for development, as 1 as for testing in each iteration.

RWG. For RWG, we adopt the same four-classifier setting as for STOP, using the word, lemma, POS, and morphological features available. For the sake of comparability with (Brunner, 2013), we train and evaluate on individual sentences, as opposed to entire documents. We use 10-fold cross validation again, randomly partitioning all corpus sentences into 10 subsets. We use GloVe embeddings pre-trained on the German Wikipedia.[1]

Evaluation. Previous studies on PARC3 adopted an exact span match setting, i.e., only those predicted spans that exactly match a gold standard span count as true positives. We report precision, recall, and F_1 in this setting for PARC3 and STOP. For RWG, we report the sentence-level accuracy used by Brunner (2013). In this mode, we train and predict with our model as before, but for evaluation we just record whether the model predicts the presence of a quotation type in a sentence.

4.2 Results

PARC3. The results in Table 1 show that NQD cannot beat the performance of Scheible et al. (2016), but does almost as well as Pareti et al. (2013). Given that our model is substantially simpler than either of these two (both include a special cue classifier, dictionaries, etc.), we see this as a success. Our model is competitive with the Scheible et al. model with regard to recall, but shows subpar precision for all quotation types, indicating a remaining weakness in the input encoding: for direct quotations, quote characters should provide strong indicators for quotation boundaries.

Note that these results, as well as the earlier studies (Pareti et al., 2013; Scheible et al., 2016), use unrealistic gold standard features. Therefore, we ran a second version of NQD using only word features, but no tags or structural information. The model is clearly worse, but still surprisingly good at 61% F_1. Not surprisingly, we see the highest drop for indirect quotations, which are most sensitive to syntactic structure. This indicates that NQD does a reasonable job in a setting that is realistic in general, and particularly so for non-standard corpus varieties (e.g., historical and literary corpora) that are often used in Digital Humanities.

STOP. To our knowledge, the results in Table 2 are the first modeling results on STOP. In comparison to PARC3, the results are noticeably lower. It is still the case that direct quotations are easiest to find, but their F_1 is somewhat lower than in PARC3. Indirect quotations are much more difficult, and free indirect quotations essentially impossible. This involves multiple factors: (a) STOP is

[1] Available from deepset at https://deepset.ai/german-word-embeddings

Model	Features	Overall			Direct			Indirect			Mixed		
		Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
Pareti et al. (2013)	word, syn, domain	63	**80**	71	88	**94**	91	56	**78**	65	60	67	63
Scheible et al. (2016)	word, syn, domain	**71**	79	**75**	93	**94**	**94**	**64**	73	**69**	68	**81**	**74**
NQD	word, syn	**71**	67	69	**94**	82	88	**64**	64	64	**70**	59	64
NQD	word	61	61	61	90	84	87	53	56	54	60	54	57

Table 1: Results on PARC3 (exact span match evaluation)

Model	Features	Overall			Direct			Indirect			Free Indirect			Reported		
		Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
NQD	word	51	66	57	78	83	80	33	49	40	01	04	01	46	58	51

Table 2: Results on STOP (exact span match evaluation)

Model	Features	Overall			Direct			Indirect			Free Indirect			Reported		
		Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
Brunner (2013) RB	word, syn	**71**	67	**69**	**87**	81	84	**62**	**81**	**71**	**44**	24	31	**64**	51	**57**
Brunner (2013) ML	word, syn	63	77	**69**	85	**88**	**87**	47	62	53	29	63	40	45	**56**	50
NQD	word, syn	60	**78**	68	77	86	82	52	69	60	31	**68**	**42**	34	**56**	43
NQD	word	59	73	65	77	83	80	40	69	50	14	62	23	41	50	45

Table 3: Results on RWG (sentence-level accuracy evaluation)

significantly smaller, but more varied, than PARC3, providing sparser training data; (b) STOP covers a wider variety of quotation types, some of these are intrinsically difficult to model – in particular free indirect quotations (McHale, 2009).

RWG. The results in Table 3 show a picture that is overall similar to PARC3:[2] NQD does not outperform the state-of-the-art, but approximates it closely despite the lack of corpus-specific tuning. As in STOP, we see the lowest results for free indirect quotations, showing that this class is generally hard to classify. In general, even though this resource's size and annotation are similar to STOP, we see significantly higher numbers. This is mostly due to the different evaluation we use for RWG to compare to previous work: detecting the presence of quotes is easier than identifying their spans.

On RWG, we also run a basic NQD with only word form information. With this corpus and evaluation, this results in a drop of merely 3 points F_1 – due to losses on the indirect and free indirect categories – which bolsters the potential of this configuration.

[2] Brunner (2013) does not report overall results. We compute them as micro-averages over reported per-type results.

5 Error Analysis

To gain some insights into the failure modes of NQD, we perform a brief qualitative analysis of the cases where our model gave false predictions.

These errors can broadly be divided into three categories: cases where the model predicts the presence of extraneous quotations (false positives), cases where the model fails to identify existing quotations (false negatives), and cases where the model correctly identified the presence of a quote, but did not correctly determine its boundaries (boundary mismatch, leading both false positives and false negatives in our exact span evaluation). We focus our error analysis on PARC, the previously best explored of our three corpora. In the examples, gold-standard quotations are marked with red square brackets, as above, and model-predicted quotations are marked with blue parentheses.

5.1 False Positives

Among the false positives produced by our model was a surprising number of quotations that are correct according to PARC's guidelines, but which are not annotated in the corpus. As an example, our model correctly identifies the presence of an

unannotated quotation in the following sentence:

(3) (Britain's retail price index rose 0.7% in September from August and was up 7.6% for the year), the Central Statistical Office said.

Outside of these cases, proper false positives seem to be rare. Many of the false positives we found were boundary mismatches, discussed separately below.

5.2 False Negatives

Among the false negatives we analyzed, we found that the model is most likely to miss "tricky" quotations that are unusual in their grammatical structure. In particular, it tends to miss a class of quotations that are expressed as short noun phrases or adjectival phrases embedded within a non-quotation sentence such as

(4) Mandela, considered [the most prominent leader of the ANC] remains in prison. But [his release within the next few months] is widely expected.

According to the PARC guidelines, these are cases of quotations since they are attributable statements, but they are difficult for the model to retrieve since they are hard to distinguish from "non-quotation" nominal phrases – in particular in cases like this one, where there are not even overly realized speakers. In STOP and RWG, these cases might arguably not even be annotated as quotations.

5.3 Boundary Mismatches

A large proportion of the errors of NQD are boundary errors, where the model identifies the presence of a quotation, but fails to identify its exact boundaries. This can happen when our model correctly predicts one quotation boundary, but not the other.

For example, in the following sentence, our classifier identified the first quotation's beginning, but not its end (it also failed to identify the second quotation entirely – a false negative):

(5) He reiterated ([his opposition to such funding], but expressed [hope of a compromise].

This type of error occurs both for noun phrases and verb phrases and embedded sentences, but for different reasons: noun phrases are difficult to recognize since they are not marked by punctuation as are almost all other cases of quotation spans; verb phrases, on the other hand, can become arbitrarily complex. In the case above, the segmentation problems are exacerbated by the fact that the noun phrase quotation span occurs in a complex syntactic environment involving coordination.

6 Conclusion

In this paper, we have argued that existing models of automatic quotation annotation suffer from the tight relation between corpus annotation and model properties in particular in terms of model reusability. As an alternative, we have presented a general neural architecture, NQD, that can be trained "as is" on various corpora that differ in terms of genre, structure, and language. While the model does not reach the state of the art on any particular corpus, it performs close to it on all of them. Notably, the model is also able to deal relatively graciously with the absence of linguistic information. We will release an implementation with pre-trained models.

As NQD makes independent predictions for each token, it cannot model correlations and mutual exclusions between labels, and there is no guarantee for well-formed output class sequences. We investigated a number of extensions, including linear-chain CRF layers that are effective for Named Entity Recognition (Lample et al., 2016), but did not obtain improvements. We believe this is due to the unbounded length of quotation spans which is challenging for CRFs (Scheible et al., 2016).

The overall greatest challenge that NQD faces is data scarcity — all existing corpora with manual annotation are small, and our results show consistently bad performance for infrequent quotation types. In this situation, transfer learning seems like a natural proposition, and our model makes it possible for the first time to apply straightforward transfer learning to quotation annotation. In future work, we will explore this direction.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of Alice in Wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada. Association for Computational Linguistics.

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation

attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48. Association for Computational Linguistics.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden.

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Manaal Faruqui and Sebastian Pado. 2012. Towards a model of formal and informal address in English. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–633, Avignon, France.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark.

Fotis Jannidis, Albin Zehe, Leonard Konle, Andreas Hotho, and Markus Krug. 2018. Analysing direct speech in German novels. In *Proceedings of DhD*, Cologne, Germany.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*, pages 260–270, San Diego, CA.

Brian McHale. 2009. Speech representation. In Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert, editors, *Handbook of Narratology*. De Gruyter.

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-character sentiment analysis in Shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Silvia Pareti. 2015. *Attribution: A Computational Approach*. Ph.D. thesis, University of Edinburgh.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, WA.

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

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492, Borovets, Bulgaria.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of ACL*, pages 1736–1745, Berlin, Germany.

Elena Semino and Michael Short. 2004. *Corpus Stylistics: Speech, Writing And Thought Presentation In A Corpus Of English Writing*. Routledge Advances In Corpus Linguistics. Routledge, London.