Detecting negation scope is easy, except when it isn't

Federico Fancellu[1] **Adam Lopez**[1] **Bonnie Webber**[1] **Hangfeng He**[2]
[1]ILCC, School of Informatics, University of Edinburgh
[2]School of Electronics Engineering and Computer Science, Peking University
{f.fancellu}@sms.ed.ac.uk, {alopez, bonnie}@inf.ed.ac.uk, hangfenghe@pku.edu.cn

Abstract

Several corpora have been annotated with *negation scope*—the set of words whose meaning is negated by a cue like the word "not"—leading to the development of classifiers that detect negation scope with high accuracy. We show that for nearly all of these corpora, this high accuracy can be attributed to a single fact: they frequently annotate negation scope as a single span of text delimited by punctuation. For negation scopes not of this form, detection accuracy is low and undersampling the easy training examples does not substantially improve accuracy. We demonstrate that this is partly an artifact of annotation guidelines, and we argue that future negation scope annotation efforts should focus on these more difficult cases.

1 Introduction

Textual *negation scope* is the largest span affected by a negation cue in a negative sentence (Morante and Daelemans, 2012).[1] For example, given the marker **not** in (1), its scope is *use the 56k conextant modem*.[2]

(1) I do **not** [use the 56k conextant modem] since I have cable access for the internet

Fancellu et al. (2016) recently presented a model that detects negation scope with state-of-the-art accuracy on the Sherlock Holmes corpus, which has been annotated for this task (SHERLOCK; Morante and Daelemans, 2012). Encoding an input sentence and cue with a bidirectional LSTM, the model predicts, *independently* for each word, whether it is in or out of the cue's scope.

But SHERLOCK is only one of several corpora annotated for negation scope, each the result of different annotation decisions and targeted to specific applications or domains. Does the same approach work equally well across all corpora? In answer to this question, we offer two contributions.

1. We evaluate Fancellu et al. (2016)'s model on all other available negation scope corpora in English and Chinese. Although we confirm that it is state-of-the-art, we show that it can be improved by making *joint* predictions for all words, incorporating an insight from Morante et al. (2008) that classifiers tend to leave gaps in what should otherwise be a continuous prediction. We accomplish this with a sequence model over the predictions.

2. We show that in all corpora except SHERLOCK, negation scope is most often delimited by punctuation. That is, in these corpora, examples like (2) outnumber those like (1).

(2) It helps activation , [**not** inhibition of ibrf1 cells] .

Our experiments demonstrate that negation scope detection is very accurate for sentences like (2) and poor for others, suggesting that most classifiers simply overfit to this feature of the data. When we attempt to mitigate this effect by undersampling examples like (2) in training, our system does not improve on examples like (1) in test, suggesting that more training data is required to make progress on the phenomena they represent. Given recent interest in improving negation annotation (e.g. Ex-Prom workshop 2016), we recommend that future negation scope annotations should fo-

[1]Traditionally, negation scope is defined on logical forms, but this definition grounds the phenomenon at word level.

[2]For all examples in this paper, negation cues are in bold, human-annotated negation scope is in square brackets [ ], and automatically predicted negation scope is underlined.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 58–63,
Valencia, Spain, April 3-7, 2017.

cus on these cases.[3]

2 Models

We use the bi-directional LSTM of Fancellu et al. (2016). The input to the network is a negative sentence w = $w_1...w_{|w|}$ containing a negation cue. If there is more than one cue, we consider each cue and its corresponding scope as a separate classification instance. Given a representation c of the cue, our model must predict a sequence $s = s_1...s_{|w|}$, where $s_i = 1$ if w_i is in the scope defined by c, and 0 otherwise. We model this as $|w|$ independent predictions determined by probability $p(s_i|w, c)$, where the dependence on w and c is modeled by encoding them using a bidirectional LSTM; for details refer to Fancellu et al. (2016).

Although this model is already state-of-the-art, it is natural to model a dependence between the predictions of adjacent tokens. For the experiments in this paper, we introduce a new joint model $p(s|w, c)$, defined as:

$$p(s|w, c) = \prod_{i=1}^{n} p(s_i|s_{i-1}, w, c)$$

The only functional change to the model of Fancellu et al. (2016) is the addition of a 4-parameter transition matrix to create the dependence on s_{i-1}, enabling the use of standard inference algorithms. This enables us to train the model end-to-end.

3 Experiments

We experiment with two English corpora: the SFU product review corpus (Konstantinova et al., 2012); and the BioScope corpus (Vincze et al., 2008). The latter consists of three subcorpora: abstracts of medical papers (ABSTRACT), full papers (FULL) and clinical reports (CLINICAL).

We also experiment with the Chinese Negation and Speculation (CNeSp) corpus (Zhou, 2015), which also consisting of three subcorpora: product reviews (PRODUCT), financial articles (FINANCIAL) and computer-related articles (SCIENTIFIC).

3.1 Corpus differences

Although they all define the scope as *the tokens in a sentence affected by a negation cue* (Morante and Daelemans, 2012), these corpora are quite different from SHERLOCK, which deals with a wider range of complex phenomena including ellipsis, long-range dependencies and affixal negation. Though widely used (e.g. Qian et al. (2016)), the SFU, BioScope and CNeSp corpora contain simplifications that are sometimes hard to justify linguistically. In SFU and BioScope, for instance, scope is usually annotated only to the right of the cue, as in (1). The only exception is passive constructions, where the subject to the left is also annotated:

(3) [This book] **wasn't** [published before the year 2000.]

On the other hand, in the CNeSp corpus, subjects are usually annotated as part of the scope, except in cases like VP-coordination (4). This is to ensure that the scope is always a continuous span.

(4) 酒店有高档的配套设施,然而却[不能多给我们提供一个枕头]
*The hotel are furnished with upscale facilities, but [***cannot** *offer us one more pillow]*

Unlike in the other corpora, in SHERLOCK, negation scope frequently consists of multiple disjoint spans of text, including material that is omitted in CNeSp. In addition to annotating the subject, as shown above, this corpus also annotates auxiliaries (5) and entire clauses (6).

(5) [...] the ground [was] damp and [the night] **in**[clement].

(6) [An investigator needs] facts and **not** [legends or rumours] .

Sherlock also annotates scope inside NPs, for example, when the the adjective bears affixal negation:

(7) I will take [an] **un**[pleasant remembrance] back to London with me tomorrow

3.2 Experimental parameters

All of our corpora are annotated for both cue and scope. Since we focus on scope detection, we use gold cues as input. We train and test on each corpus separately. We first extract only those sentences containing at least one negation cue (18% and 52% for English and Chinese respectively) and create a 70%/15%/15% split of these for training, development and test respectively. We use a fixed split in order to define a fixed development set for error analysis, but this setup

[3] http://www.cse.unt.edu/exprom2016/

precludes direct comparison to most prior work, since, except for Fancellu et al. (2016), most has used 10-fold cross-validation. Nevertheless, we felt a data analysis was crucial to understanding these systems, and we wanted a clear distinction between test (for reporting results) and development (for analysis).

Model parameters and initialization are the same as in Fancellu et al. (2016). We pretrain our Chinese word embeddings on wikipedia and segment using NLPIR.[4,5] For Chinese, we experimented with both word and character representations but found no significant difference in results.

Baseline. In preliminary experiments, we noticed many sentences where negation scope was a single span delimited by punctuation, as in (2). To assess how important this feature is, we implemented a simple baseline in three lines of python code: we mark the scope as all tokens to the left or right of the cue up until the first punctuation marker or sentence boundary.

3.3 Results

We evaluate our classifier in two ways. First, we compute the *percentage of correct scopes* (PCS), the proportion of negation scopes that we *fully* and *exactly* match in the test corpus. Second, we measure token-level F_1 over tokens identified as within scope. To understand the importance of continuous spans in scope detection, we also report the number of gaps in predicted scopes.

Results are shown in Table 1, including those on SHERLOCK for comparison.[6] It is clear that the LSTM system improves from joint prediction, mainly by predicting more continuous spans, though it performs poorly on CNeSp-SCIENTIFIC, which we believe is due to the small size of the corpus. More intriguingly, the baseline results clearly demonstrate that punctuation alone identifies scope in the majority of cases for SFU, BioScope, and CNeSp.

[4]Data from `https://dumps.wikimedia.org/`

[5]NLPIR: `https://github.com/NLPIR-team/NLPIR`

[6]Unlike all other corpora where the scope if always continuous and where the joint prediction helps to ensure no gaps are present, in *Sherlock* the gold scope is often discontinuous; this is the reason why we also cannot test for gaps.

Data	System	F_1	PCS	gaps
Sherlock	Baseline	68.31	26.20	-
	Fancellu et al. (2016)	**88.72**	63.87	-
	+joint	87.93	**68.93**	-
SFU	Baseline	87.07	77.90	-
	Cruz et al. (2015)*	84.07	58.69	-
	Fancellu et al. (2016)	**89.83**	74.85	17
	+joint	88.34	**78.09**	**0**
BioScope Abstract	Baseline	82.75	64.59	-
	Zou et al. (2013)*	-	76.90	-
	Fancellu et al. (2016)	91.35	73.72	37
	+joint	**92.11**	**81.38**	**4**
BioScope Full	Baseline	75.30	50.41	-
	Velldal et al. (2012)*	-	**70.21**	-
	Fancellu et al. (2016)	**77.85**	51.24	20
	+joint	77.73	54.54	**6**
BioScope Clinical	Baseline	97.76	94.73	-
	Velldal et al. (2012)*	-	90.74	-
	Fancellu et al. (2016)	97.66	**95.78**	4
	+joint	**97.94**	94.21	**1**
CNeSp Abstract	Baseline	81.70	70.57	-
	Zhou (2015)*	-	60.93	-
	Fancellu et al. (2016)	90.13	67.35	26
	+joint	**90.58**	**71.94**	**0**
CNeSp Financial	Baseline	90.84	58.87	-
	Zhou (2015)*	-	56.07	-
	Fancellu et al. (2016)	**94.88**	**75.05**	6
	+joint	93.58	74.03	**0**
CNeSp Scientific	Baseline	83.43	31.81	-
	Zhou (2015)*	-	**62.16**	-
	Fancellu et al. (2016)	**81.30**	40.90	4
	+joint	80.90	59.09	**0**

Table 1: Results for the English corpora (Sherlock, SFU & BioScope) and for Chinese corpora (CNeSp). * denotes results provided for context that are not directly comparable due to use 10-fold cross validation, which gives a small advantage in training data size.

Data	Punctuation	Other
Sherlock	68%	45%
SFU	92%	23%
BioScope Abstract	88%	51%
BioScope Full	84%	30%
BioScope Clinical	98%	47%
CNeSp Product	80%	37%
CNeSp Financial	84%	66%
CNeSp Scientific	20%	41%
Total	85%	40%
Average	85%	40%

Table 2: PCS results on the development set, split into cases where punctuation exactly delimits negation scope in the gold annotation, and those where it does not.

4 Error analysis

The baseline results suggest that punctuation alone is a strong predictor of negation scope, so we further analyze this on the development set by dividing the negation instances into those whose scopes (in the human annotations) are precisely delimited by the innermost pair of punctuation markers containing the cue, and those which are not. The results (Table 2) confirm a huge gap in accuracy between these two cases. The model correctly learns to associate surrounding punctuation with scope boundaries, but when this is not sufficient, it underpredicts, as in (8), or overpredicts, as in (9).

(8) surprisingly , expression of [**neither** bhrf1 **nor** blc-2 in a b-cell line , bjab , protected by the cells from anti-fas-mediated apostosis] ...

(9) ..., 下次是肯定[不 会再住锦地星座了]

*Next time (I) [won**'t** live again in Pingdi Xingzuo] for sure*

A closer inspection reveals that in SHERLOCK, where this gap is narrower, we correctly detect a greater absolute number of the difficult punctuation scopes, though accuracy for these is still lower. The results on CNESP- SCIENTIFIC may again be due to the small corpus size.

To understand why the system is so much better on punctuation-delimited scope, we examined the training data to see how frequent this pattern is (Table 3). The results suggest that our model may simply be learning that punctuation is highly indicative of scope boundaries, since this is empirically true in the data; the fact that the SHERLOCK and CNESP-SCIENTIFIC are the exception to this is in line with the observations above.

This result is important but seems to have been overlooked: previous work in this area has rarely analyzed the contribution of each feature to classification accuracy. This applies to older CRF models (e.g. Morante et al. (2008)), as well as to more recent neural architectures (e.g. CNN, Qian et al. (2016)), where local window based features were used.

In order to see whether training imbalance was at play, we experimented with training by undersampling from training examples that can be predicted by scope boundaries only. We report results on using incrementally bigger samples of the majority class. Figure 1 shows the results for the SFU corpus, which is a representative of a trend we observed in all of the other corpora. There does indeed seem to be a slight effect where the classifier overfits to punctuation as delimiter of negation scope, but in general, classification of the other cases improves only slightly from under-sampling. This suggests that the absolute number of training instances for these cases is insufficient, rather than their ratio.

Data	Total	Punctuation
Sherlock	984	40%
SFU	2450	80%
BioScope Abstract	1190	64%
BioScope Full	210	54%
BioScope Clinical	560	93%
CNeSp Product	2744	71%
CNeSp Financial	1053	58%
CNeSp Scientific	109	22%

Table 3: Training instances by corpus, showing total count and percentages whose scope is predictable by punctuation boundaries only.

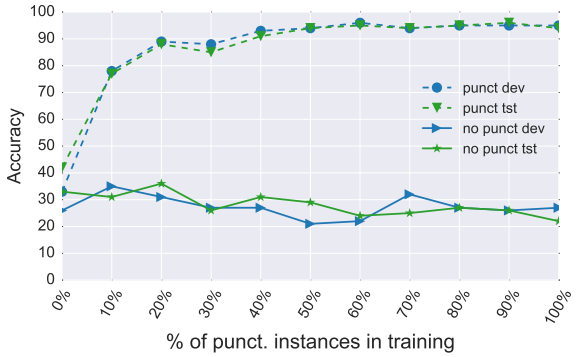

Figure 1: PCS accuracy on development and test sets divided into instances where the punctuation and scope boundaries coincide (*punct.*) and instances where they do not (*no punct.*), when *punct.* instances are incrementally removed from the training data.

5 Re-annotation of negation scope

At this point it is worth asking: is negation scope detection easy because most of the instances in real data are easy? Or is it because the annotation guidelines made it easy? Or is it because of the domain of the data? To answer these ques-

tions we conducted a small experiment on SFU, BioScope-abstract and CNeSp-financial, each representing a different domain. For each, we randomly selected 100 sentences and annotated scope following the Sherlock guidelines. If the guidelines are indeed responsible for making scope detection easy, we should observe relatively fewer instances predictable by punctuation alone in these new annotations. If instead, easy instances still outnumber more difficult ones, we can conclude that detecting negation scope is less easy on Sherlock Holmes because of the domain of the data. Comparing the results in Table 4 with the one in Table 3, the Sherlock-style annotation produces more scopes that are not predictable by punctuation boundaries than those that are. We attribute this to the fact that by capturing elliptical constructions, the Sherlock guidelines require the annotation of complex, discontinuous scopes, as in (10).

(10)

BIOSCOPE : second , t cells , which lack cd45 and **can not** [signal via the tcr] , supported higher levels of viral replication and gene expression .

BIOSCOPE-SHERLOCK : second , [t cells] , which lack cd45 and **can not** [signal via the tcr] , supported higher levels of viral replication and gene expression .

In contrast with the original SFU and BioScope annotation, always annotating the subject produces negation scopes that are not bound by punctuation, since in both English and Chinese, subjects generally appear to the left of the cue and are less often delimited by any punctuation (11).

(11)

SFU : i 'm sure she felt rather uncomfortable having to ask us at all , but she thought it was strange that we 'd **not** [mentioned it] .

SFU-SHERLOCK :i 'm sure she felt rather uncomfortable having to ask us at all , but she thought it was strange that [we 'd] **not** [mentioned it] .

Data	Punct.	No Punct.
SFU	42%	58%
BioScope Abstract	34%	64%
CNeSp Financial	45%	55%

Table 4: Percentages of scope instances predictable (punct.) and not predictable (no punct.) by punctuation boundaries only on 100 randomly selected sentences annotated following the *Sherlock* guidelines for each of the three corpora considered.

6 Discussion and Recommendation

We have demonstrated that in most corpora used to train negation scope detection systems, scope boundaries frequently correspond to punctuation tokens. The main consequence of this is in the interpretation of the results: although neural network-based sequence classifiers are highly accurate quantitatively, this appears to be so because they are simply picking up on easier cases that are detectable from punctuation boundaries. Accuracy on difficult cases not delimited by punctuation is poor. Under-sampling easy training instances seems to have little effect.

For future research in this area we make two strong recommendations. (1) Our *data-oriented* recommendation is to adopt a more linguistically-motivated annotation of negation, such as the one used in the SHERLOCK annotation, and to focus annotation on the more difficult cases. (2) Our *model-oriented* recommendation is to explore more recursive neural models that are less sensitive to linear word-order effects such as punctuation.

Acknowledgments

This project was also founded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL).

The authors would like to thank Sameer Bansal, Nikolay Bogoychev, Marco Damonte, Sorcha Gilroy, Joana Ribeiro, Naomi Saphra, Clara Vania for the valuable suggestions and the three anonymous reviewers for their comments.

References

Noa P Cruz, Maite Taboada, and Ruslan Mitkov. 2015. A machine-learning approach to negation and spec-

ulation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 495–504.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Maña López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *LREC*, pages 3190–3195.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul*. Citeseer.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 715–724. Association for Computational Linguistics.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. 2016. Speculation and negation scope detection via convolutional neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 815–825.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational linguistics*, 38(2):369–410.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1.

Bowei Zou Qiaoming Zhu Guodong Zhou. 2015. Negation and speculation identification in chinese language. In *Proceeding of the Annual ACL Conference 2015*.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2013. Tree kernel-based negation and speculation scope detection with structured syntactic parse features. In *EMNLP*, pages 968–976.