Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish

Anastassia Shaitarova
Department of Computational Linguistics
University of Zurich, Switzerland
`anastassia.shaitarova@uzh.ch`

Fabio Rinaldi
Dalle Molle Institute for Artificial
Intelligence Research, Switzerland
`fabio.rinaldi@idsia.ch`

Abstract

Negation is a linguistic universal that poses difficulties for cognitive and computational processing. Despite many advances in text analytics, negation resolution remains an acute and continuously researched question in Natural Language Processing. Reliable negation parsing affects results in biomedical text mining, sentiment analysis, machine translation, and many other fields. The availability of multilingual pre-trained general representation models makes it possible to experiment with negation detection in languages that lack annotated data. In this work we test the performance of two state-of-the-art contextual representation models, Multilingual BERT and XLM-RoBERTa. We resolve negation scope by conducting zero-shot transfer between English, Spanish, French, and Russian. Our best result amounts to a token-level F1-score of 86.86% from Spanish to Russian. We correlate these results with a linguistic negation typology and lexical capacity of the models.

1 Introduction

Negation continues to occupy the minds of many researchers. It is a fascinating and complicated linguistic phenomenon that is still not entirely understood or conceptualized. Moreover negation is an important thought process. The ability to negate is a deeply human trait that is also universal, therefore any given language is bound to have negation (Horn, 2001).

Negation has the power to change the truth value of a proposition. Thus its identification in text is of utmost importance for the reliability of results since negated information should either be discarded or presented separately from the facts. This is particularly relevant for biomedical text mining and sentiment analysis but is also important for most Natural Language Processing (NLP) tasks. The identification of negated textual spans, however, is far from trivial. Negation exhibits great diversity in its syntactic and morphological representation.

Like many other NLP tasks, most work on negation detection has been done on the English language, though there is a growing amount of research on negation detection in Spanish, Chinese and some other languages. Despite the need for quality text analytics around the world, annotated data is still sparse in many languages. This motivates the further exploration of approaches like transfer learning where models are trained on available resources and subsequently tested on a different target language.

In this paper we use a cross-lingual transfer-learning approach for negation scope detection using two state-of-the-art general purpose representation models: mBERT (Multilingual BERT, Devlin, 2018) and XLM-R (XLM-RoBERTa, Conneau et al., 2020). We fine-tune the models on freely accessible annotated corpora in English, Spanish, and French and test them cross-lingually. Additionally we test the models on a small dataset in Russian which was specially annotated for the experiment. Our research is guided by three objectives:

• We compare the performance of two state-of-the-art models on the task of cross-lingual zero-shot negation scope resolution in Spanish, French, and Russian;

• We experiment with Russian which is an under-sourced and under-researched language regarding the task of negation detection;

• We study the four involved languages typologically and correlate our findings with the experiment results.

In Section 2 we perform a brief typological analysis of the languages in relation to negation. Additionally, we overview previous work on cross-

lingual negation scope resolution. Section 3 discusses the datasets and highlights their annotation differences. We describe the experiments and present the results in Section 4, and in Sections 5 and 6 we discuss the results and draw conclusions.

2 Negation and its processing

Linguistics and typology. A number of psycholinguistic studies show that humans require extra time in order to process negation during language comprehension (Gulgowski and Błaszczak, 2020). This is attributed to the fact that humans first construct a positive counterpart of the argument and only then embed its negative aspect as an extra step (Tian and Breheny, 2016). Indeed, negative sentences exhibit a more complicated, marked-up structure on a lexico-syntactic level which is a universal feature (Barigou et al., 2018). The main building blocks of this markup are negative words and expressions, also known as negative markers, cues, or triggers.

When a sentence contains more than one negation trigger, Negative Concord (NC) languages treat them as one, letting relevant negative markers intensify one another. The majority of languages including French, Spanish, and Russian belong to the NC group. Standard English, on the other hand, is a Double Negation (DN) language where each negative marker is interpreted separately.

Hossain et al. (2020) compared English to a number of languages in regards to negation features drawn from the World Atlas of Language Structures (WALS)[1]. They showed that the number of negation-related errors in machine translation corresponds to how close the languages are in a typology based on negation.

Inspired by their discoveries we construct a negation-based typology for our languages and merge it with the classification from Dahl (1979). Even though English, French, Spanish, and Russian are in the same linguistic family and feature the same subject-verb-object pattern, the typology based on negation assigns them to different categories (Table 1).

We expect a negation-based linguistic typology to help us predict and interpret our results. According to our classification, Russian is most similar to Spanish and least similar to English. Thus we hypothesise that zero-shot transfer from Spanish into Russian will be most successful.

Lang	predNeg	symm	NC/DN	Dahl et al.
RU	yes	symm	NC	S11 12
ES	mixed	symm	NC	S11 12
FR	no	symm	NC	S11^2 12/22
EN	no	both	DN	S11/S3 22

Table 1: Negation-based typology of languages. **Pred-Neg** indicates whether negative indefinite pronouns require an additional negative particle. Symmetricity of negation (**symm**) shows whether the presence of a negation marker causes grammatical changes in the sentence. **NC/DN** means Negative Concord vs. Double Negation. In **Dahl**'s typology S11 represents a class of languages where an uninflected particle must be added while the finite verb does not change. S11^2 signals the use of double particles. Number 12 categorizes languages where a negative marker immediately precedes a finite element (verb) whereas 22 indicates that the marker immediately follows it. S3 shows the use of noninflected markers together with dummy auxiliaries.

Automated negation detection consists of two tasks: identification of negation cues, and detection of sentence parts that are affected by these cues. The latter is called *negation scope resolution*, the task that interests us most.

Negation detection began in the medical domain with the goal of improving information retrieval from Electronic Health Records (EHRs). Rule-based algorithms such as NegExpander (Aronow et al., 1999), NegFinder (Mutalik et al., 2001), NegEx (Chapman et al., 2001), and their adaptations were used in order to find medical concepts and then determine whether they are negated. The scope of negation was often understood as a distance between a negation cue and a medical term that it affects.

These algorithms were successful and some are still in wide use due to their explainability, customizability, and independence from annotated data. NegEx is incorporated into various modern computational libraries[2] and is successfully used for biomedical texts (Cotik et al., 2016; Elazhary, 2017). Despite the aforementioned qualities, rule-based algorithms suffer from an inherent inability to generalize (Wu et al., 2014; Sergeeva et al., 2019; Sykes et al., 2020).

The release of the BioScope corpus (Szarvas

et al., 2008; Vincze et al., 2008) became a pivotal moment for negation detection by providing data for machine learning. Negation scope resolution was formalized by Morante et al. (2008); Morante and Daelemans (2009), who established it as a problem of sequence classification. Using gold-standard cues and an ensemble of three different classifiers, they achieved the best F1-score of 84.71% on the Full Papers subcorpus of BioScope and 90.67% on the Abstracts subcorpus. The latter result was later surpassed by Fancellu et al. (2017) who employed neural networks and reached a score of 92.11%.

The Shared Task on Resolving the Scope and Focus of Negation (Morante and Blanco, 2012) addressed the issue of negation scope resolution directly and released another annotated corpus (ConanDoyle-neg, Morante and Daelemans, 2012). The best system (Packard et al., 2014) used an enhanced hybrid model by Read et al. (2012) and a semantic parser. They reached an F1-score of 88.2% using gold-standard cues. These results were surpassed by Li and Lu (2018) who used the Conditional Random Fields classifier and reached an F1-score of 89.4%.

Additionally, Fancellu et al. (2016) secured an F1-score of 89.93% on the SFU Review-NEG corpus (Konstantinova et al., 2012), another publicly available corpus annotated for negation scope. The results on these three corpora remained the benchmark for negation scope resolution until the Bidirectional Encoder Representation from Transformers (BERT, Devlin et al., 2019) became the new state of the art. Moreover, BERT became widely used for transfer learning due to its enhanced ability to generalize using attention and general purpose language representations. NegBERT (Khandelwal and Sawant, 2020) set new records for negation scope resolution on all three publicly available corpora.

Cross-lingual negation scope work. Many languages remain under-researched regarding negation detection and particularly scope resolution. One of the main problems is the lack of annotated data. There currently exist a handful of corpora in English, two in Spanish, and one corpus each in Swedish, German, Dutch, Chinese, Italian, and Portuguese which are not all publicly available (Jiménez-Zafra et al., 2020).

Negation work on Spanish has been growing in recent years but it has mostly concerned senti-

ment analysis (Brooke et al., 2009; Vilares et al., 2013; Jimenez Zafra et al., 2019). Rivera Zavala and Martinez (2020) are the first ones to work with sense embeddings to detect negation cues and scopes in the Spanish biomedical and general domain texts. They also worked with mBERT but in a monolingual setting. The research on negation in French is particularly limited. Aside from a few papers describing rule-based approaches (Deléger and Grouin, 2012; Abdaoui et al., 2017) and the implementation of BiLSTMs by Dalloux et al. (2019, 2020), there is barely any other research available on the topic.

Cross-lingual work on negation detection is even more limited. Fancellu et al. (2018) developed a truly cross-lingual system that uses no language specific features. They worked with English and Chinese and used universal dependencies to abstract away from the word order. Their Bidirectional Dependency LSTM model reached an F1-score of 72.46%.

Finally, Shaitarova et al. (2020) employed Multilingual BERT to perform zero-shot transfer for negation scope resolution and showed good preliminary results. We build on this work and compare mBERT with a new multilingual general purpose representation model, XLM-R. Unlike mBERT, XLM-R was pre-trained on more than two terabytes of filtered data collected by CommonCrawl. Instead of WordPiece units it uses SentencePiece (Kudo and Richardson, 2018) units and features a bigger size of shared vocabulary (250K).

3 Data

In our experiments we work with a corpus of clinical texts in French (Dalloux et al., 2020), and SFU ReviewSP-NEG, a Spanish corpus of online reviews (Jiménez-Zafra et al., 2018). The English data includes the biological paper abstracts and full scientific articles in the domain of bioinformatics from BioScope (Vincze et al., 2008), all available subcorpora of the ConanDoyle-neg corpus (Morante and Daelemans, 2012) as well as SFU (SFU Review-NEG, Konstantinova et al., 2012), a large multi-domain corpus of product reviews.

We use the English corpora separately and also combine them into one training dataset. The three corpora belong to different domains and feature certain variations in scope annotation guidelines. Despite these significant problems we combine the datasets based on the successful cross-corpora

knowledge transfer described by Khandelwal and Sawant (2020).

The BioScope annotators set the precedent by ultimately basing scope annotation on syntax. They employed a maximal scope size strategy and extended annotation to the biggest syntactic unit possible. The normal direction of scope was assumed to be to the right of the cue. The subject is not included in the scope, unless the sentence has a passive voice.

Morante et al. (2011) argued that semantically the subject should be always annotated within the scope. Thus, unlike the BioScope corpus, ConanDoyle-neg includes the subject yet excludes the cue. Additionally, it features morphological negations. The SFU corpus mostly adheres to the BioScope's annotation guidelines but does not include cues into the scope of negation.

The French data is described in Dalloux et al. (2020) and is publicly available on request[3]. It combines two subcorpora of clinical narratives. Its format and annotations are loosely modeled on the ConanDoyle-neg corpus. The data in the Spanish SFU ReviewSP-NEG corpus can be requested from Simon Fraser University. Its annotations reflect the guidelines of the English corpora but are also based on Spanish grammar.

In our experiments we only use sentences that contain at least one negation. We duplicate sentences with multiple negations into several copies containing a single negation. Table 2 shows the statistics for all the corpora. For the sake of consistency we excluded cues from scope annotation across all corpora.

	ConDo	BioScope	SFU	SP	FR
uniq	1215	1935	3112	3258	1682
negs+	1421	2095	3528	4327	1870

Table 2: Corpora statistics. `uniq` indicates the original number of unique sentences with negations. `neg+` shows the number of negation sentences after the duplication of sentences with multiple negations.

3.1 The Russian test set

To the best of our knowledge, there are no publicly available negation corpora in Russian or any other Slavic language. Thus, there is almost no available research on negation detection in Russian

on either the English or Russian speaking Internet, with Funkner et al. (2020) being the only relevant publication.

In order to work with Russian in our experiments, we created a small dataset annotated with negation cues and negation scopes[4]. It is a Russian counterpart to one of the ConanDoyle-neg's test sets, The Adventure of the Cardboard Box. The number of sentences containing negation amounts to 120.

The annotation was performed by one native Russian speaker using Prodigy[5], an annotation tool created by explosion.ai. Since there are no known publications about negation detection for Russian, the annotation was based on linguistic intuition, Russian grammar, and a generalization of annotation schema from the other corpora.

In accordance with the guidelines, the scope in the Russian test set corresponds to a syntactic component. A maximal scope rule was implemented as in BioScope. The subject is included in the scope when the negation cue directly affects the main verb. Cues are not included in the scope. Since morphological cues appear only in ConanDoyle-neg, they were not considered during annotation.
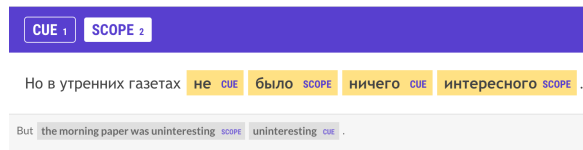

Figure 1: Annotation of negation cues and scopes in a Russian sentence with the use of the Prodigy annotation tool.

4 Experiments and results

We used NegBERT (Khandelwal and Sawant, 2020) as the main architecture and employed `bert-base-multilingual-cased` and `xlm-roberta-base-model` models. We fine-tuned the two models on the three datasets: English (**en**), Spanish (**es**), and French (**fr**). All the models were trained with the same set of hyperparameters. Early stopping method with patience set to 9 was used to prevent overfitting. The maximum input length was adjusted to 250 to prevent truncated sentences.

The word-level token class is determined by using the argmax function on the averaged softmax

---

[3]http://people.irisa.fr/Clement.Dalloux/

[4]https://doi.org/10.5281/zenodo.4537833

[5](Montani and Honnibal, 2018)

probabilities of all subword units. We use gold-standard negation cues and report token-level F1-scores for negation scope resolution (Table 3).

Despite the fact that the English corpora are of different domains, models fine-tuned on the combined English data brought better cross-lingual results than models that were fine-tuned on each corpus individually. Even the model fine-tuned on the ConanDoyle-neg corpus did not perform better on the Russian version of the text. Thus, we only discuss the results of the model trained on the entirety of the English data.

Since the datasets differ in size, we ran additional experiments where we equalized the number of training examples to the smallest corpus (French). We drew a random sample of 1870 sentences from the English and the Spanish data and retrained the models. Row ru2 in Table 3 shows these evaluation results.

	EN		FR		ES	
fr	82.61	83.22	–	–	79.33	79.79
es	78.62	78.83	76.81	78.17	–	–
ru	83.47	86.49	76.50	80.07	81.40	86.73
ru2	80.07	85.35	76.50	80.07	81.24	86.86

Table 3: Evaluation results for mBERT (grey columns) and XLM_RoBERTa (white columns). The models were fine-tuned on English (**EN**), French (**FR**), and Spanish (**ES**) and tested on French (**fr**), Spanish (**sp**), and Russian (**ru**). Row **ru2** shows evaluations of models that were fine-tuned on equal size data.

5 Discussion and error analysis

There have been many debates on whether BERT-like models truly "understand" negation. Zhao and Bethard (2020) showed evidence for shallow encoding of this phenomenon in both BERT and RoBERTa. Meanwhile Staliūnaitė and Iacobacci (2020) demonstrated that these models lack linguistic abstraction abilities and fail when confronted with compositional semantic aspects of language.

In our experiments, the XLM-R model performed significantly better than mBERT for all language pairs. As an additional metric, we measured how well both models identified scopes with 100% precision. Averaged across all languages, both models performed equally well, with mBERT solving 46.23% of exact scopes, and XLM-R - 46.66%. The best result for Russian was produced by the XLM-R model fine-tuned on Spanish (53% of exact scopes).

In fact, Russian benefited most from a transfer from Spanish and least from French, irrespective of training data size or model type. We can assume that the success of the Spanish-Russian transfer is partially due to the commonalities described in Table 1. Nevertheless, the negation typology does not explain the poor results of the French-Russian pair.

We investigated several factors that might have negatively affected the French-Russian knowledge transfer. For example, we examined the vocabularies of the models and calculated lexical overlap between the datasets based on a model-specific tokenization. The comparison in Table 4 shows a lower percentage of lexical overlap between the Russian and the French datasets than between Russian and other languages. According to this observation, however, English-Russian transfer should have been the most successful one.

	vocab size		shared vocab					
	SubW	SentP	**en**		**fr**		**es**	
en	10550	10592	–	–	23	21	28	26
es	8453	8934	35	31	22	17	–	–
fr	5101	5032	47	44	–	–	36	31
ru	1280	1329	35	31	23	21	28	26

Table 4: Vocabulary distribution across the data. Numbers in grey are calculated on the basis of mBERT's SubWord units. White columns show XLM-R's SentencePiece units. Numbers in the shared vocab section indicate percentages. For example, the French data shares 47% of its WordPieces and 44% of its SentencePieces with the English dataset, while only 23% of the English SubWords and 21% of its SentencePieces appear in the French dataset.

Next, we took a closer look at our negation typology. We investigated a prominent phenomenon that emerges in several categories, namely negative indefinite pronouns (words like *nothing, nowhere, nobody*). The way a languages handles these pronouns is reflected in both the **predNeg** and the **NC/DN** columns in Table 1. This phenomenon classifies Russian and English as polar opposites.

We found 19 sentences in the Russian dataset that contain negation structures with negative indefinite pronouns. Despite the fact that these pronouns are always marked as cues, the English XLM-R model included them into the scope 9 times. The English mBERT model made that same mistake 3 times. On the other hand, neither Spanish, nor French models had this problem. We can hypothe-

sise that a model fine-tuned on English could not coordinate a negative particle with an indefinite negative pronoun in the same sentence since this does not occur in English.

During the examination of these 19 sentences we stated that the models fine-tuned on the French data persistently omit the subject of a sentence in the annotation of scope. The English models also suffered from this problem but to a lesser extent. This can be traced to the difference in annotation. The subject is not annotated in the French corpus while only part of the English data features that annotation. Figure 2 illustrates the issue of negative indefinite pronouns as well the annotation of a sentence's subject.

ES: Нет, [я] **ничего** [такого] **не** [видел].
EN: Нет, я [**ничего** такого] **не** [видел].
FR: Нет, я **ничего** [такого] **не** [видел].
      No, I nothing such not saw.

      No, [I saw] **nothing**.

Figure 2: A Russian sentence with a negative indefinite pronoun featuring annotations by three XLM-R models fine-tuned on Spanish (ES), English (EN), and French(FR). The fourth line contains a literal translation. The bottom line is the original sentence and annotation from the ConanDoyle-neg corpus.

Additionally we investigated scope annotations which were precisely identified by one type of model but not the other. We chose to look at the highest scoring language pair Spanish-Russian where the models were trained on 1870 sentences. There are 15 sentences where the XLM-R model found scope with a perfect precision while mBERT did not. In most cases mBERT made a mistake in the leftward direction from a negation cue.

We found only four cases where mBERT scored perfectly while XLM-R made a mistake. The mistakes are rather random and do not seem to belong to a particular pattern. Overall, we detected several situations where mistakes made by the models could be scrutinized due to questionable annotation. We acknowledge that the lack of additional annotators and an inter-annotator agreement is a weakness that should be addressed in further work.

6 Conclusion

The short excursion into negation scope resolution in Russian using zero-shot model transfer has shown good preliminary results. Despite contro-

versial previous findings, multilingual general purpose representation models perform rather well on negation scope resolution. XLM-RoBERTa scored consistently better than mBERT in all language pairs.

We constructed a typology that classifies English, Spanish, French, and Russian according to their negation-based features. Since indefinite negative pronouns play a role in several typological categories, we investigated their effect on zero-shot transfer. We found that fine-tuning models on English compromises their performance with this phenomenon when transferring to Russian, which correlates with the negation typology.

Transferring syntactic negation knowledge from Spanish brought the most benefit for Russian. This result is fully in line with the negation typology of the four languages. Despite the clear correlation between the negation typology and the results of the Spanish-Russian transfer, not all outcomes are easily explainable. The relatively poor performance of the French-Russian transfer might be related to the domain mismatch and the difference in annotation schemes. A lower lexical overlap between the vocabularies could have had an effect as well.

Future work involves growing the Russian corpus of negations, ideally benefiting from multiple annotators. It may prove beneficial to perform a systematic examination of all the categories constituting the negation typology and to expose their effects on knowledge transfer across languages.

References

Amine Abdaoui, Andon Tchechmedjiev, William Digan, Sandra Bringay, and Clement Jonquet. 2017. French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français.

David B. Aronow, Feng Fangfang, and W. Bruce Croft. 1999. Ad Hoc Classification of Radiology Reports. *Journal of the American Medical Informatics Association : JAMIA*, 6(5):393–411.

Baya Naouel Barigou, Fatiha Barigou, and Baghdad Atmani. 2018. Handling Negation to Improve Information Retrieval from French Clinical Reports. *Journal of e-Learning and Knowledge Society*, 14(1). Number: 1.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria. Association for Computational Linguistics.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. ArXiv: 1911.02116.

Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. 2016. Negation Detection in Clinical Reports Written in German. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 115–124, Osaka, Japan. The COLING 2016 Organizing Committee.

Östen Dahl. 1979. Typology of sentence negation. *Linguistics*, 17(1-2):79–106. Publisher: De Gruyter Mouton Section: Linguistics.

Clément Dalloux, Vincent Claveau, and Natalia Grabar. 2019. Speculation and Negation detection in French biomedical corpora. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 223–232, Varna, Bulgaria. INCOMA Ltd.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2020. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, pages 1–21. Publisher: Cambridge University Press.

Louise Deléger and Cyril Grouin. 2012. Detecting Negation of Medical Problems in French Clinical Notes. In *Proc of Int Health Inform*, Miami Beach, FL.

Jacob Devlin. 2018. Multilingual BERT Readme document. Library Catalog: github.com.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Hanan Elazhary. 2017. NegMiner: An Automated Tool for Mining Negations from Electronic Narrative Medical Documents.

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: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, and Bonnie L. Webber. 2018. Neural Networks for Cross-lingual Negation Scope Detection. *ArXiv*, abs/1810.02156.

Anastasia Funkner, Ksenia Balabaeva, and Sergey Kovalchuk. 2020. Negation Detection for Clinical Text Mining in Russian. *arXiv:2004.04980 [cs]*. ArXiv: 2004.04980 version: 1.

Piotr Gulgowski and Joanna Błaszczak. 2020. Psycholinguistic Investigation of the Immediate Interpretation of Plural Nouns in the Scope of Sentential Negation in Polish. *Journal of Psycholinguistic Research*, 49(5):741–760.

Laurence R. Horn. 2001. *A natural history of negation*. The David Hume series. CSLI, Stanford, Calif.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a Non-Issue: Negation as a Source of Error in Machine Translation. *arXiv:2010.05432 [cs]*. ArXiv: 2010.05432.

Salud Maria Jimenez Zafra, M. Teresa Martin Valdivia, Eugenio Martinez Camara, and L Alfonso Urena Lopez. 2019. Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. *IEEE Transactions on Affective Computing*, 10(1):129–141.

Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, M. Dolores Molina-González, and L. Alfonso Ureña-López. 2018. Relevance of the SFU ReviewSP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish. *Information Processing & Management*, 54(2):240–251.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora Annotated with Negation: An Overview. In *Computational Linguistics*, volume 0, pages 1–87.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution. In *LREC*.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hao Li and Wei Lu. 2018. Learning with Structured Representations for Negation Scope Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 533–539, Melbourne, Australia. Association for Computational Linguistics.

I Montani and M Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. Learning the Scope of Hedge Cues in Biomedical Texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36, Boulder, Colorado. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. page 6.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the Scope of Negation in Biomedical Texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 715–724, Honolulu, Hawaii. Association for Computational Linguistics.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Pradeep Mutalik, Aniruddha M. Deshpande, and Prakash M. Nadkarni. 2001. Research Paper: Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *JAMIA*.

Woodley Packard, Emily M. Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan. 2014. Simple Negation Scope Resolution through Deep Parsing: A Semantic Solution to a Semantic Problem. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO1: Constituent-Based Discriminative Ranking for Negation Resolution. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 310–318, Montréal, Canada. Association for Computational Linguistics.

Renzo Rivera Zavala and Paloma Martinez. 2020. The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study. *JMIR Medical Informatics*, 8(12):e18953.

Elena Sergeeva, Henghui Zhu, Peter Prinsen, and Amir Tahmasebi. 2019. Negation Scope Detection in Clinical Notes and Scientific Abstracts: A Feature-enriched LSTM-based Approach. *AMIA Summits on Translational Science Proceedings*, 2019:212–221.

Anastassia Shaitarova, Lenz Furrer, and Fabio Rinaldi. 2020. Cross-lingual transfer-learning approach to negation scope resolution. In *CEUR Workshop Proceedings*, Zurich. CEUR-WS. ISSN: 1613-0073.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and Lexical Semantics in RoBERTa, BERT and DistilBERT: A Case Study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7046–7056, Online. Association for Computational Linguistics.

D. Sykes, A. Grivas, C. Grover, R. Tobin, C. Sudlow, W. Whiteley, A. Mcintosh, H. Whalley, and B. Alex. 2020. Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, pages 1–22.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Ye Tian and Richard Breheny. 2016. Dynamic Pragmatic View of Negation Processing. In Pierre Larrivée and Chungmin Lee, editors, *Negation and Polarity: Experimental Perspectives*, Language, Cognition, and Mind, pages 21–43. Springer International Publishing, Cham.

David Vilares, Miguel Ángel Alonso, and Carlos Gómez-Rodríguez. 2013. Supervised polarity classification of Spanish tweets based on linguistic knowledge. In *Proceedings of the 2013 ACM symposium*

on Document engineering, DocEng '13, pages 169–172, Florence, Italy. Association for Computing Machinery.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The Bio-Scope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11):S9.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLOS ONE*, 9(11):1–11.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.