Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, Deborah Ribeiro Carvalho

Natural Language Engineering (2020), 1–21
doi:10.1017/S1351324920000352

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora

Clément Dalloux[1,*], Vincent Claveau[1], Natalia Grabar[2], Lucas Emanuel Silva Oliveira[3], Claudia Maria Cabral Moro[3], Yohan Bonescki Gumiel[3] and Deborah Ribeiro Carvalho[3]

[1]Univ Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes, France, [2]UMR 8163 STL CNRS, Université de Lille, 59000 Lille, France and [3]Pontifícia Universidade Católica do Paraná (PUC-PR), R. Imac. Conceição, 1155 - Prado Velho, Curitiba - PR, 80215-901, Brazil
*Corresponding author. E-mail: clement.dalloux@irisa.fr

(Received 22 May 2019; revised 27 March 2020; accepted 15 May 2020)

Abstract

Automatic detection of negated content is often a prerequisite in information extraction systems in various domains. In the biomedical domain especially, this task is important because negation plays an important role. In this work, two main contributions are proposed. First, we work with languages which have been poorly addressed up to now: Brazilian Portuguese and French. Thus, we developed new corpora for these two languages which have been manually annotated for marking up the negation cues and their scope. Second, we propose automatic methods based on supervised machine learning approaches for the automatic detection of negation marks and of their scopes. The methods show to be robust in both languages (Brazilian Portuguese and French) and in cross-domain (general and biomedical languages) contexts. The approach is also validated on English data from the state of the art: it yields very good results and outperforms other existing approaches. Besides, the application is accessible and usable online. We assume that, through these issues (new annotated corpora, application accessible online, and cross-domain robustness), the reproducibility of the results and the robustness of the NLP applications will be augmented.

Keywords: Corpus annotation; Machine learning; Natural language processing for biomedical texts; Information extraction

1. Introduction

Detecting negation in texts is one of the unavoidable prerequisites in many information retrieval and extraction tasks. In the biomedical field in particular, negation is very common and plays an important role. In the case of cohort selection for clinical trials, for instance, it can provide decisive criteria for recruiting a patient or not. Thus, it provides crucial information in many situations such as: detecting a patient's pathologies and co-morbidities, determining a person's smoking or non-smoking status, detecting whether or not a particular medication has been prescribed or taken, and defining whether a patient is pregnant or not at the time of recruitment. In order to efficiently identify negation instances, one must first identify the negation cues, that is, words (or morphological units) that express negation, and secondly identify their scopes, that is, tokens within the sentence which are affected by these negation cues.

Figure 1 provides a general overview of our contribution and of its presentation. We first present the specificity of expressing negation in French and Brazilian Portuguese: the issues that can arise with examples from our corpora are described in Section 2. Some of the existing work

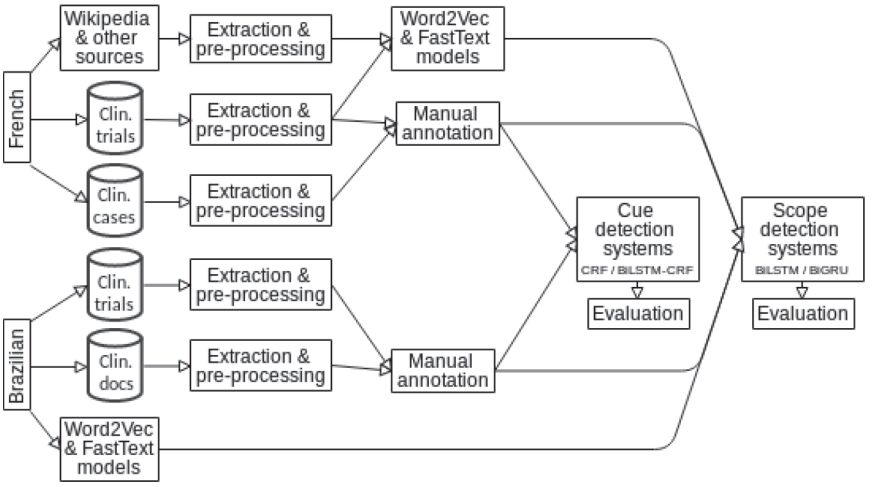

Figure 1. General overview of our work: from the development of annotated corpora in French and Brazilian Portuguese to the automatic detection of negation and its scope.

in this field is presented in Section 3. In Section 4, we describe our manually annotated corpora, which provide the training and evaluation material for the approach. The proposed approach for negation detection, with all its components (word vector representations, recurrent neural network, conditional random fields (CRF), and the evaluation rationale), is then detailed in Section 5. The results obtained are presented and discussed in Sections 6 and 7 for the cue and scope detection, respectively. Finally, Section 8 provides conclusive remarks and introduces some directions for future work.

2. Expression of negation in French and Brazilian Portuguese

As pointed out in the literature (Chapman *et al.* 2001; Elkin *et al.* 2005; Denny and Peterson 2007; Gindl, Kaiser and Miksch 2008), negation is frequently used and plays an important role in the biomedical field. However, its expression is very variable and thus it presents a first challenge for its automatic detection. As stated earlier, cue detection can be a rather complex task, due in part to the variety and ambiguity of negations marks. Moreover, extracting the scope of these negation is necessary to decide which part of the sentence is negated. Some specific linguistic realizations of negation in the two languages under consideration, French and Brazilian Portuguese, are introduced hereafter.

2.1 Negation in French

In French, the negation cues either consist of one word/prefix or of multiple words. Moreover, negation can be expressed via a large panel of cues which can be morphological, such as the following prefixes *an, in, im, ir, dis*; lexical, such as *absence de (absence of)*, *à l'exception de (except)*; and grammatical, such as *non, ne. . .pas, ni. . .ni*. In the following examples, we present and explain sentences with instances of negation which either correspond to specific situations in the detection of negation scope or are proper to the biomedical language (in the rest of the paper, the cues are underlined, scopes in bold, scope altering tokens in brackets).

1. *En alternative des traitements locaux (chirurgie, radiothérapie, radiofréquence, cryoablation) peuvent être indiqués mais ils ne* ***sont*** *pas* ***[toujours]*** *faisables. (Alternatively, local treatments*

(surgery, radiotherapy, radiofrequency, cryoablation) may be indicated but are not ***[always]*** *feasible.)*

2. *Il n'****existe toujours pas aujourd'hui de consensus quant à une définition précise de ce phénomène hétérogène ou des modalités de sa prise en charge****. (There is* ***still no consensus today on a precise definition of this heterogeneous phenomenon or the modalities of its management****.)*
3. *il n'y a pas de traitement curateur de la maladie [en dehors de] l'allogreffe de moelle. (there is no curative treatment for this disease [apart from] bone-marrow homograft)*
4. *Autre immunothérapie concomitante, excepté* ***les corticostéroïdes à faible dose****. (Other concomitant immunotherapy, except* ***low dose corticosteroids****.)*
5. *[Lymphome non hodgkinien] à cellules B matures récidivant/réfractaire. (Relapsed/refractory mature B-cell [non-Hodgkin's lymphoma].)*
6. *Cancers bronchiques [non à petites cellules]. ([Non-small cell] bronchial cancers.)*
7. *Le traitement par tazemetostat continuera jusqu'à progression de la maladie ou l'apparition d'un* ***[effet indésirable]*** *inacceptable. (The treatment with tazemetostat will be maintained until progression of the disease or appearance of an unacceptable [adverse effect].)*
8. *Elle n'est soulagée que par la marche et doit donc écouter la télévision en faisant les cent [pas] dans son salon. (She is only relieved by walking and must therefore listen to the TV [pacing] in her living room.)*

Examples 1 and 2 show the possible effect of the frequency adverbs, here *toujours (always)*, on negation. In Example 1, *traitements locaux (chirurgie, radiothérapie, radiofréquence, cryoablation)*, the content would be negated without *toujours (always)*. In Example 2, with or without *toujours*, the meaning of the sentence does not change, therefore, the scope of the negation remains the same.

Example 3 shows how the preposition, *en dehors de (apart from)*, can stop the scope of negation. Many other prepositions such as *à part*, *à l'exception de* or *excepté*, with more or less the same meaning than *en dehors de (apart from)*, would have the same effect on the negation scope. However, these prepositions can also play the role of negation by themselves, as shown in Example 4.

Examples 5–7 show that cues can also be included in medical concepts such as ***non*** *hodgkinien (non-Hodgkin's)*, ***non*** *à petites cellules (non-small cell)*, or *effet* ***in****désirable (adverse effect)*. In biomedical texts, such sequences correspond to single concepts such as identified by single UMLS CUIs. Hence, we do not consider the consider the sub-concept negation: *Hodgkin's* and *small cell* are not negated.

Finally, Example 8 shows the context in which the ambiguous word *pas* (meaning both *no/not* and *footstep*) does not bear the meaning of a negation. Indeed, in this example, ***pas*** is part of the idiomatic expression ***faire les cent pas*** *(pacing, walking around)*. This is the real issue for cue detection, yet, fortunately, the latter form only appears twice in our data. Another example of ambiguity of the kind is related to the adverb ***plus*** meaning either ***more*** or, in conjunction with ***ne***, ***no more***.

2.2 Brazilian Portuguese negation

In Brazilian Portuguese, the main negation cues are lexical units, such as *sem (without)*, *nega (denies)*, *não (no)*, or *ausência de (absence of)*. In some cases, the cues may also correspond to prefixes like *in, im, des, dis,* or *a*. Below, we provide some typical examples of negation found in clinical narratives and clinical protocols.

1. *Crise de dor pontada em hipocôndrio direito. Não* ***relacionada com a alimentação****. (Crisis of puncture pain in the right hypochondrium. Not related to food.)*

2. *Ausência de* ***irritação peritoneal****. (Absence of peritoneal irritation.)*
3. *Nega* ***outras comorbidades****. (Denies other comorbidities.)*
4. *Nega* ***dispareunia****,* ***corrimento*** *e* ***cólica****. (Denies dyspareunia, vaginal discharge and colic.)*
5. *Relata muito des****conforto*** *mesmo usando a cinta diariamente. (Reports a lot of discomfort even using the tape daily.)*
6. *Lesão no menisco in****dolor*** *no momento. (Meniscus injury that is painless at the moment.)*
7. *Exame Físico: bom estado geral, hidratada e a****febril****. (Physical Examination: good general condition, hydrated and afebrile.)*
8. *Sem* ***lesões*** *ou* ***fratura de ossos****. (No injury or bone fracture.)*
9. *Declarar não* ***estar gestante****. (Declare not being pregnant.)*
10. ***Teste de elevação da perna estendida*** *negativo. (Negative extended leg lift test.)*

Examples 1–7 are provided by the corpus of clinical narratives. The most traditional and frequent cue is the grammatical cue *não* (*not*) as presented in Example 1. In Example 2, the negation is based on the use of the lexical cue *absence of*. As one can also see, the scope of one negation cues may concern a single (Example 3) or multiple entities when they are coordinated (Example 4). In Examples 5–7, the negation is stated by the use of prefixes.

Examples 8–10 are provided by a corpus of clinical trials recruitment protocols. In these protocols, the third person verbs and pronouns are never used: hence, the writing style is different from the one used in narratives. Consequently, some of the ways of expressing the negation are specific: uses of negative non-verbal sentences (Example 8), the verb *declare* is often used to assess having or not certain conditions (Example 9), mentions of negative results for a given condition, examination or lab results (Example 10).

3. Related work

Negation detection is a very well researched problem. In this section, we present several corpora and methods that have been proposed in the literature.

3.1 Data

In the recent years, with the democratization of supervised machine learning techniques, several specialized corpora in English have been annotated with negation-related information, which has resulted in pre-trained models for automatic detection. These corpora can be divided into two categories: (1) corpora annotated with cues and scopes of negation, such as BioScope and *SEM-2012, and (2) corpora focusing on medical concepts/entities, such as i2b2 and MiPACQ. We briefly describe these corpora. The BioScope corpus (Vincze *et al.* 2008) contains reports of radiological examinations, scientific articles as well as abstracts from biomedical articles. Available in the XML format, each sentence and each negation and uncertainty cue/scope pair receives a unique identifier. Table 1 provides some statistics about each subcorpus in BioScope. We can see for instance that the prevalence of sentences with negation and uncertainty is high. The *SEM-2012 corpus (Morante, Schrauwen and Daelemans 2011) consists of a Sherlock Holmes novel and three other short stories written by Sir Arthur Conan Doyle. It contains 5520 sentences, among which 1227 sentences are negated. Each occurrence of the negation, the cue and its scope are annotated, as well as the focus of the negation if relevant. The peculiarity of this corpus is that cues and scopes can be discontinuous, as indicated in the annotation guidelines. In addition to lexical features, that is, lemmas, the corpus also offers syntactic features, that is, part-of-speech tagging and chunking. The i2b2/VA-2010 challenge (Uzuner *et al.* 2011) featured several information retrieval and extraction tasks using US clinical records. One of the tasks involved the detection of assertions, that is, each medical problem concept (diseases, symptoms, etc.) was associated with one of six assertion types: *present*, *absent*, *possible*, *conditional*, *hypothetical,* or *not associated with the*

Table 1. Statistics of the BioScope corpora (Vincze *et al.* 2008)

	Examinations	Articles	Abstracts
Documents	1954	9	1273
Sentences	6383	2670	11,871
Negative sentences	13.55 %	12.70 %	13.45 %
Negation cues	877	389	1848
Uncertain sentences	13.39 %	19.44 %	17.70 %
Uncertainty cues	1189	714	2769

patient. MiPACQ (Albright *et al.* 2013) is another corpus which consists of clinical data in English annotated with several layers of syntactic and semantic labels. Each detected UMLS entity has two attribute locations: *negation*, which can take two values (*true* or *false*), and *status*, which can take four values (*none*, *possible*, *HistoryOf,* or *FamilyHistoryOf*).

3.2 Rule-based systems

Among the rule-based systems dedicated to negation detection, *NegEx* (Chapman *et al.* 2001), which pioneered the area, is still popular. This system uses regular expressions to detect the cues and to identify medical terms in their scope. Later this system was adapted to other languages such as Swedish (Velupillai, Dalianis and Kvist 2011) and French (Deléger and Grouin 2012). *NegFinder* (Mutalik, Deshpande and Nadkarni 2001) is another pioneering system which combines a lexical analyzer, which uses regular expressions to generate a finite state machine, and a parser, which relies on a restricted subset of the non-contextual Look-Ahead Left Recursive grammar. *NegFinder* makes it possible to identify the concepts impacted by negation in medical texts when they are close to the linguistic units marking the negation. Derived from *NegEx*, *ConText* (Harkema *et al.* 2009) covers additional objectives. This system detects negation, temporality, as well as the subject concerned by this information in the clinical texts. It has been adapted to French (Abdaoui *et al.* 2017). *NegBio* (Peng *et al.* 2017) relies on rules defined from the universal dependency graph (UDG). The code for this system is available online.[a]

3.3 Supervised machine learning

The system of Velldal *et al.* (2012) considers the set of negation cues as a closed class. This system uses an SVM and simple n-grams features, calculated on the words and lemmas, to the right and left of the candidate cues. This system offers a hybrid detection of the scope of negation. It combines expert rules, operating on syntactic dependency trees, with a ranking SVM, which operates on syntagmatic constituents. It was further improved by Read *et al.* (2012) and is used as a fallback by Packard *et al.* (2014) when the main MRS (minimal recursion semantics) Crawler cannot parse the sentence. Fancellu, Lopez and Webber (2016) use neural networks to solve the problem of negation scope detection. One approach uses feed-forward neural network, an artificial neural network where the connections between the units do not form loops. Another approach, which appears to be more efficient for the task, uses a bidirectional Long Short-Term Memory (biLSTM) neural network. Li and Lu (2018) use CRF, semi-Markov CRF, as well as latent-variable CRF models to capture negation scopes.

[a] https://github.com/ncbi-nlp/NegBio, last accessed in March, 9th 2020.

Table 2. Statistics on the four corpora created and annotated

	French clin. trials	French clin. cases	Brazilian clin. trials	Brazilian clin. narratives
Documents	644	200	285	1000
Sentences	6547	3811	3228	9808
Tokens	150,084	87,487	48,204	156,166
Vocabulary (types)	7880	10,500	6453	15,127
Negative sentences	1025	804	643	1751
IAA (cue tokens)	0.9001	0.9933	–	0.7414
IAA (scope tokens)	0.8089	0.8461	–	

Résumé
L'objectif de cet essai est d'évaluer l'efficacité de la détection du papillomavirus humain (HPV) par la méthode d'auto-prélèvement, chez des patientes ne participant pas au dépistage du cancer du col de l'utérus. Les patientes seront recrutées lors d'une consultation gynécologique habituelle. Au cours de cette consultation, trois prélèvements de différents types seront réalisés : deux seront des auto-prélèvements, réalisés par la patiente (coton tige), le troisième sera réalisé par le médecin (frottis cervico-utérin standard). Les patientes recevront par courrier les résultats des tests de détection du HPV réalisés sur les trois prélèvements.

Figure 2. Example of the summary from a clinical trial protocol in French.

4. Data: Creation of annotated corpora in French and Brazilian

There are very few, if any, corpora which are annotated with negation-related information in languages other than English. In order to train effective machine learning-based models for negation detection in the languages we work with, we developed our own annotated corpora. The corpora that we describe in this section were developed in cooperation by French and Brazilian researchers. Table 2 presents some statistics on these corpora: the number of words, the variety of the vocabulary, the number of sentences, the number of sentences with one or more negations, and the inter-annotator agreements (IAA). However, as explained in Section 4.5, we cannot provide any IAA for the Brazilian clinical protocols. The IAA provided for the Brazilian clinical narratives was computed on both cue and scope tokens.

4.1 ESSAI: French corpus with clinical trial protocols

Our first corpus contains clinical trial protocols in French. The clinical trial protocols are mainly obtained from the National Cancer Institute registry[b]. As shown in Figures 2 and 3, the typical French protocol consists of two parts. First, the summary of the trial indicates the purpose of the trial and the applied methods. Then, another document, from a different page of the website, describes in detail the trial, and particularly, the inclusion and exclusion criteria. In both those parts, negation provides useful information regarding the specification of the target cohort and the recruitment of patients. As shown in Table 2, the ESSAI corpus is our second largest corpus in terms or sentences and tokens and contains more than a thousand sentences with at least one instance of negation (about 16% of all sentences).

[b] https://www.e-cancer.fr

APACHE-1 : Essai comparant la détection du papillomavirus humain par auto-prélèvement vaginal à l'examen standard par frottis cervico-utérin, chez des patientes ne participant pas au dépistage cytologique du cancer du col de l'utérus. [essai clos aux inclusions]
Résumé scientifique / schéma thérapeutique
Il s'agit d'un essai de dépistage non randomisé et multicentrique.
Les patientes sont recrutées dans le cadre habituel d'une consultation gynécologique. Les patientes ont 3 prélèvements vaginaux. Deux sont des auto-prélèvements, l'un en milieu sec et l'autre en milieu liquide. Le 3ème prélèvement (frottis cervico-utérin) est réalisé par le médecin au cours de l'examen gynécologique.
Les patientes sont informées par courrier des résultats des tests HPV, réalisés en aveugle sur les 3 prélèvements.
Objectif principal
Valider une approche par auto-prélèvement vaginal pour la détection d'infections cervicales à papillomavirus humains oncogènes.
Objectif secondaire
Comparer la performance du test HPV sur auto-prélèvement vaginal transporté soit en milieu liquide, soit sans milieu liquide, Comparer la distribution des génotypes d'HPV sur les prélèvements vaginaux et cervicaux.
Critères d'inclusion
Age $\geq$ 20 ans et $\leq$ 65 ans, Consentement éclairé signé.
Critères de non-inclusion
Menstruation, Vaccination contre le HPV, Virginité, Hystérectomie totale, Pathologie cervicale HPV dépendante en cours de traitement, Frottis anormal de moins d'un an, Frottis normale de moins de deux ans, Femme ne comprenant pas l'objectif de l'étude ou la langue, Femme enceinte.
Critère d'évaluation principal
Présence d'une infection à HPV selon les différents prélèvements.

Figure 3. Example of the detailed description from clinical trial protocol in French.

4.2 CAS: French Corpus with Clinical Cases

The CAS corpus (Grabar, Claveau and Dalloux 2018) contains clinical cases in French. The collected clinical cases are issued from different journals and websites from French-speaking countries (e.g., France, Belgium, Switzerland, Canada, African countries. . .). These clinical cases are related to various medical specialties (e.g., cardiology, urology, oncology, obstetrics, pulmonology, gastroenterology). The purpose of clinical cases is to describe clinical situations for real (de-identified) or fake patients. Common clinical cases are typically part of education programs used for the training of medical students, while rare cases are usually shared through scientific publications for the illustration of less common clinical situations. As for the clinical cases found in legal sources, they usually report on situations which became complicated due to various reasons: medical doctor, healthcare team, institution, health system, and their interactions. Similarly to clinical documents, the content of clinical cases depends not only on the clinical situations illustrated and on the disorders but also on the purpose of the presented cases: description of diagnoses, treatments or procedures, evolution, family history, expected audience, etc. Figure 4 shows a typical example of a clinical case from the CAS corpus. The document starts with the introduction of the patient (44-year-old female), explaining why she was hospitalized (1). The following sentences (2) add some information about the patient's history of Crohn's disease. Then, the next section (3) describes the examination and lab results obtained for this patient. Finally, the last sentence (4) describes which treatment was chosen, its effect on the patient, and the outcome of the healthcare process. In clinical cases, the negation is frequently used for describing patient signs and symptoms and for the diagnosis of patients. It can also be used for the description of the patient evolution. As shown in Table 2, this negation annotated version of the CAS corpus currently contains fewer sentences than ESSAI. However, CAS has a larger vocabulary and percentage of sentences with at least one instance of negation (about 21% of all sentences).

(1) Une femme de 44 ans était hospitalisée en juillet 1999 pour une diarrhée évoluant depuis la veille, faite de 8 selles diurnes et 3 selles nocturnes, glairo-sanglantes et impérieuses. La diarrhée était associée à des douleurs hypogastriques. [...]

(2) Dans ses antécédents, la malade avait eu deux résections iléales pour une maladie de Crohn sténosante de l'intestin grêle diagnostiquée en 1977. La dernière intervention avait été réalisée en octobre 1998 pour un abcès de l'anse iléale pré-anastomotique. [...]

(3) L'examen clinique à l'admission était sans particularité, hormis une douleur provoquée à la palpation de la fosse iliaque gauche et de l'hypogastre. Les examens biologiques montraient une hyperleucocytose à 11,4 G/L et un syndrome inflammatoire avec une vitesse de sédimentation à 50 mm à la première heure. L'hémoglobinémie était normale à 14 g/dL. [...]

(4) En 24 heures, sous repos digestif et après arrêt des antibiotiques, l'évolution était favorable avec un retour à l'état clinique antérieur à la colite aiguë.

Figure 4. Example of the clinical case in French.

Título Público:
A influência da palmilha proprioceptiva associada a acupuntura em mulheres.
Título Científico:
A influência da palmilha proprioceptiva associada à acupuntura sobre o equilíbrio, postura, atividade muscular, flexibilidade e perfil energético dos meridianos em mulheres.
Texto:
O grupo G1 (n=15) utilizaram palmilhas comum para sapato e receberá a informação para utilizar a palmilha comum por 4 horas por dia
O grupo G2 (n=15) e G3 (n=15) receberá a palmilha confeccionada individualmente, após avaliação postural e clínica e terão que utilizá-las por 4 horas por dia [...]
Os grupos serão avaliados na pré, imediatamente após, 30, 60 e 90 dias após a intervenção.
O número total de indivíduos para os grupos serão 45, sendo 15 para cada grupo.
Critério de Inclusão:
Estudantes universitárias; sexo feminino; idade entre 18-30 anos; sedentárias;
ter acesso ao Whatsapp; aplicativo para smartphones; afim de para facilitar a à comunicação.
Critério de Exclusão:
Estudantes que apresentam escoliose em S; alterações vestibulares; auditiva e oculares não corrigida; fraturas recentes em membro inferior; prótese no membro inferior; gestante; IMC > 25.
Design do Estudo:
Ensaio clinico tratamento, randomizado-controlado, paralelo, aberto, com dois braços.

Figure 5. Example of a clinical protocol in Brazilian Portuguese.

4.3 Brazilian clinical trial protocols

The Brazilian clinical trial protocols were provided by the dedicated Brazilian website[c]. Figure 5 shows an excerpt from Brazilian clinical trial protocol. Its structure is similar to the French example presented above. Each protocol indicates its public title, scientific title, a description of the trial, as well as the inclusion and exclusion criteria. In these documents, negation provides useful information regarding the specification of the target cohort and the recruitment of patients. As shown in Table 2, this corpus has fewer sentences, tokens, and negative sentences than the others (about 20% of all sentences).

[c] http://ensaiosclinicos.gov.br/, last accessed in March, 9th, 2020.

S: ACOMPANHA NO AMBULATORIO POR PROLAPSO GENITAL DE SEGUNDO GRAU, COM RETENÇAO URINARIA NEGA INCONTINENCIA,TROUXE PRE OPERATORIOS E AVALIACAO LIBERADA PELO CARDIOLOGISTA [...]
O: UREIA 43 TGP 27 TGO 18 PU NORMAL
ECG 01/11/13: FC 69 SINUSAL
USTV: UETRO RVF HETEROGENIO DOIS NNODULOS INTRAMURAIS 1,4X1,7X1,8CM E OUTRO DE 0,8X0,4X0,6 CM VOL 33 CM 3 ENDOMETRIO 2,8 MM OVARIOS NORMAIS [...]
P: LIBERO AIH E SOLICITO NOVOS EXAMES LABORATORIAIS E CARTA PRO ANESTESISTA RETORNO EM 04/04/14

Figure 6. Example of a clinical narrative in Brazilian Portuguese.

4.4 Brazilian clinical narratives

The clinical narratives were provided by three Brazilian hospitals and are related to several medical specialties such as cardiology, nephrology, or endocrinology. Thousand documents of various nature (discharge summaries, medical nursing notes, ambulatory records, clinical evolution, etc.) were manually annotated with negation cues and their scope. An example of Brazilian Portuguese clinical narrative is presented in Figure 6. This example follows the SOAP (Subjective, Objective, Assessment and Plan) note structure, which corresponds to the typical organization of clinical notes created by the healthcare workers. As shown in Table 2, this is our largest corpus in terms of sentences, tokens, vocabulary, and negative sentences (about 18% of all sentences).

4.5 Annotation process

In this section, we present the basic annotation guidelines that were used to develop all four corpora; we also present the annotation process and its results. Regarding negation cues, annotators were asked to annotated any token usually triggering negation (if several cues are present in a sentence, they are numbered) that is not part of a biomedical concept identified by a French or Portuguese UMLS CUI. Accordingly, *Cancer bronchique non à petites cellules* (C0007131) is not annotated; however, *mésothéliome pleural malin non résécable* is annotated (1). Regarding negation scopes, we only annotated the sequences of tokens which contained the focus of the negation (with the identifying number of the corresponding cue if needed). As our purpose is to create data and methods usable in a clinical environment, we only annotate a medical concept as part of a scope when it is the focus of the negation instance. For instance, in Example 1, *mésothéliome pleural malin* should not be marked as negated since the sentence cannot be rephrased as *no malignant pleural mesothelioma*. Conversely, in Example 2, *une autopsie* is part of the scope as we can summarize the sentence as *no autopsy*. Example 3 shows that if the effect of the negation cue is altered by another token, the annotated scope is affected. Indeed, without *toujours*, *Le traitement de référence* would be part of the scope.

1. Mésothéliome pleural malin non **résécable**. *(Unresectable malignant pleural mesothelioma.)*
2. **Une autopsie** n'**était** pas **réalisée**. *(An autopsy was not performed.)*
3. Le traitement de référence est chirurgical mais celui-ci n'**est** pas **toujours possible**. *(The reference treatment is surgical, but is not always possible.)*

The annotators were given a description of the task and examples as guidelines. The annotation of both French corpora involved three annotators. One annotated both corpora and the others each annotated either ESSAI or CAS. While the ESSAI corpus was annotated manually to mark up negation cues and scopes, the CAS corpus was first annotated automatically with models trained on the ESSAI corpus. Then, the two annotators manually verified every single sentence in order to correct annotations and annotate forgotten instances of negation. Regarding cue annotation, the resulting IAAs are strong, a Cohen's kappa coefficient of 0.9001 on the ESSAI corpus and of 0.9933 on the CAS corpus. While disagreements on the CAS corpus were a few forgotten

Table 3. Excerpts from the two French corpora. The columns contain linguistic information (lemmas, POS-tag), negation cues, and their scope

Sentence	Position	Form	Lemma	POS Tag	Cue	scope
4937	0	Homme	homme	NOM	_	_
4937	1	40	@card@	NUM	_	_
4937	2	ans	an	NOM	_	_
4937	3	sans	sans	PRP	sans	_
4937	4	ATCD	ATCD	NAM	_	ATCD
4937	5	.	.	SENT	_	_
3146	0	Sem	sem	ADP	sem	_
3146	1	comprometimento	compromet	NOUN	_	comprometimento
3146	2	da	da	X	_	da
3146	3	capacidade	capac	NOUN	_	capacidade
3146	4	funcional	funcional	ADJ	_	funcional
3146	5	.	.	PUNCT	_	_

cues during the manual annotation process, most of the disagreements on the ESSAI corpus were related to biomedical concepts. Indeed, biomedical concepts such as *anti-angiogéniques* and *anti-tumorale* were annotated by one annotator. During the adjudication process, it was decided that those would not be annotated. From these annotations, we can compute the most frequent cues: *ne. pas* (53 % of all cues), *non* (18 %), *sans* (13 %), *aucun* (7 %). Regarding scope annotation, the resulting IAAs are strong as well. A Cohen's kappa coefficient of 0.8089 on the ESSAI corpus and of 0.8461 on the CAS corpus. Apart from the scopes associated with cue disagreements, only a few tokens were annotated differently on the ESSAI corpus. For instance, *this* from the sequence *this has not been demonstrated in immunocompromised patients* was wrongly annotated as part of the scope. Disagreements on the CAS corpus were mostly related to punctuation and subjects annotated as part of the scope, for instance, *the patient* from the sequence *the patient had no clinical recurrence*.

The annotation of the Brazilian clinical protocols involved three students from the Pontifical Catholic University of Paraná. However, they all ended up annotating different parts of the corpus for lack of time; therefore, IAA are unavailable. Regarding Brazilian clinical narratives, seven students and a nurse took part in the annotation process. The resulting level of agreement between annotators is rather high, with a Cohen's kappa coefficient of 0.7414. Ultimately, the nurse, along with a physician, took part in the adjudication process. The annotation process is described in detail in Oliveira *et al.* (2020). The most frequent cues are *naõ* (39 %), *sem* (11 %), *in-* (9 %), *de-* (6 %), and *ausência* (5 %).

Moreover, Table 3 presents one annotated sentence in each language in the *CoNLL* format: *40-year-old male without clinical history* and *without compromising functional capacity*. All corpora offer several additional annotation layers. The French corpora were pre-processed with TreeTagger (Schmid 1994) for part-of-speech tagging and lemmatization. The Brazilian part-of-speech tags (*Universal POS tags*) were obtained using *RDRPOSTagger* (Nguyen *et al.* 2014). The stems were obtained using the Portuguese Snowball Stemmer from NLTK[d].

[d]https://www.nltk.org/

5. Methodology

The second purpose of our work is to design a cross-domain approach for the automatic and effective detection of negation (cues and their scope). In this section, we describe the methods that were designed and tested. They rely on specifically trained word vectors and supervised learning techniques. The objective is to predict whether each word is a part of the negation cue and/or scope or not.

5.1 Word vector representations

Various methods have been used to represent words as vectors. Let's mention for instance, the bag-of-word models like simple token counts or TF-IDF, to which Latent Dirichlet allocation (Blei, Ng and Jordan 2003) or Latent semantic analysis (Deerwester *et al.* 1990) can be applied. Even though these approaches are still used, several more recent models have been introduced for a better representation of semantic relations between words needed by machine learning approaches. We present the methods we use for training the word vectors prior to the negation detection task.

word2vec (Mikolov *et al.* 2013) is a particularly efficient predictive model to learn word embeddings from plain text. Word embeddings can be calculated using two model architectures: the continuous bag-of-words (CBOW) and *Skip-Gram* (SG) models. Algorithmically, these models are similar except that CBOW predicts the target words from the words of the source context, while the *skip-gram* model does the opposite and predicts the source context words from the target words.

fastText (Bojanowski *et al.* 2017) addresses the word2vec's main issue: the words which do not occur in the vocabulary cannot be represented. Indeed, word2vec ignores the morphological structure of words and only assigns the features based on their semantic context. Hence, the authors address this limitation by using subword information: each word is represented as a bag of all possible character n-grams it contains. The word is padded using a set of unique symbols which helps singling out prefixes and suffixes. With a large enough corpus, every possible character n-gram may be covered, and, since the representations across words are often shared, rare words can also get reliable representations.

In French, the two word embedding models are trained using the *Skip-Gram* algorithm, 100 dimensions, a window of five words left and right, a minimum count of five occurrences for each word, and negative sampling. The training data are composed of the French Wikipedia articles and biomedical data. The latter includes the ESSAI and CAS corpora, the French Medical Corpus from CRTT,[e] and the Corpus QUAERO Médical du français[f] (Névéol *et al.* 2014). These models are trained using the Gensim[g] (Rehurek and Sojka 2010) python library. In Brazilian Portuguese, we use the pre-trained models available on the *NILC* (*Núcleo Interinstitucional de Lingustica Computacional*) website[h] (Hartmann *et al.* 2017). These models were obtained using the *Skip-Gram* algorithm and 100 dimensions. In English, we use two fastText models. The original fastText model: 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens); and our own model trained on Conan Doyle's novels, the assumption being that this domain-specific model will outperform models trained on generic data. The latter was trained with Gensim using CBOW, hierarchical softmax, and 100 dimensions for 100 epochs. For all languages, we also use randomly initialized vectors as input to our neural networks. In this case, the weights are initialized very close to zero, but randomly. Lemma and part-of-speech embeddings are randomly initialized as well. Then, those embeddings' weights are updated during training.

[e] https://bit.ly/2LOJfEW
[f] https://quaerofrenchmed.limsi.fr/
[g] https://radimrehurek.com/gensim/
[h] http://nilc.icmc.usp.br/embeddings

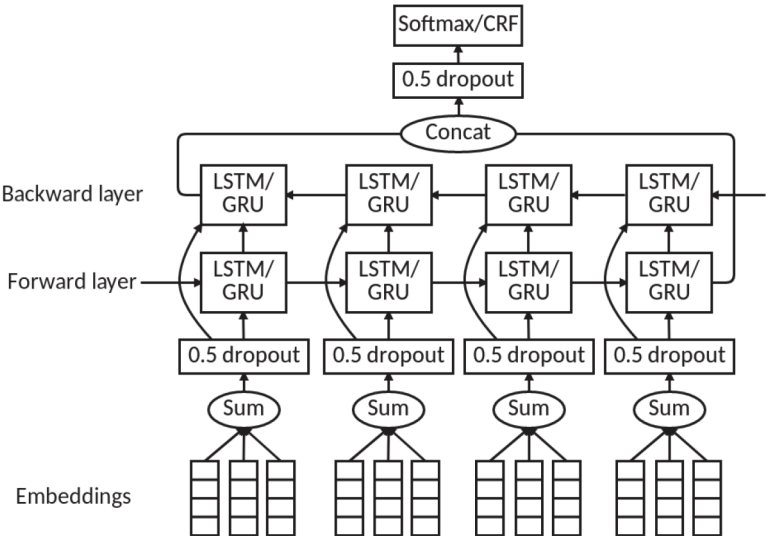

Figure 7. Our BiRNN uses LSTM or GRU cells and a softmax or CRF output layer.

5.2 Recurrent neural network

A recurrent neural network is a class of network which is capable of adapting its decision by taking into account the previously seen data, in addition to the currently seen data. This operation is implemented thanks to the loops in the architecture of the network, which allows the information to persist in memory. Among the RNNs, long short-term memory networks (LSTM) (Hochreiter and Schmidhuber 1997) are the most efficient at learning long-term dependencies and are therefore more suitable to solve the problem of discontinuous scope, which is typical with negation. The gated recurrent unit (GRU) network (Cho *et al.* 2014) is a variant of the LSTM where the forget and input gates are merged into one single update gate. The cell state C and the hidden state h are also merged. The created model is therefore simpler than the model obtained with the standard LSTM.

In our experiments, we used a bidirectional recurrent neural network, which operates forward and backward, to detect both negation cues and scopes. The backward pass is particularly relevant for the scope detection because the negated words may occur before or after the cue. Implemented with *Keras* using *TensorFlow* (Abadi *et al.* 2016) as backend, our systems include French-language and Brazilian Portuguese-language versions of the bidirectional LSTM inspired by Fancellu *et al.* (2016), as well as a bidirectional GRU (BiGRU). Prediction is computed either by a *softmax* layer, which is the most common method, or by a CRF layer, which seems to be particularly suitable for sequence labeling. An overview of the BiRNN architecture is shown in Figure 7. We use embedding layers of dimension $k = 100$ with 0.5 dropout and a dimensionality of the output space of 400 units per layer (backward/forward) with 0.5 recurrent dropout. Fifty epochs seem more than enough to achieve the highest possible F_1 score on the validation sets.

5.3 Conditional random fields

CRFs (Lafferty, McCallum and Pereira 2001) are statistical methods used in natural language processing to label word sequences. CRFs generally obtain good results with much lower training time than neural networks. In our experiments, we performed gradient descent using the L-BFGS (Limited-memory BFGS) method. We only experiment with CRFs for the cue detection task, in comparison with the BiLSTM-CRF model.

Table 4. Results for the cue detection task on the four corpora. The results are given as Precision, Recall and F_1-score. The best scores are in bold

System	Corpus	Window size	P	R	F_1
CRF	French protocols	(4)	96.05	91.89	93.92
BiLSTM-CRF		None	99.09	93.70	**96.32**
CRF	French cases	(4)	97.05	97.37	97.21
BiLSTM-CRF		None	96.99	98.17	**97.58**
CRF	Brazilian protocols	(2)	90.67	86.08	88.31
BiLSTM-CRF		None	90.73	86.71	**88.67**
CRF	Brazilian narratives	(3)	88.60	90.41	89.49
BiLSTM-CRF		None	94.64	90.71	**92.63**

5.4 Evaluating labeling systems

To evaluate our systems, we use the standard evaluation measures: precision P, which quantifies the relevance of the automatic categorization, recall R, which quantifies the sensitivity of the automatic categorization, as well as F_1-score (harmonic mean of the precision and recall noted F_1). To evaluate the detection of the negation scope, we compute those measures in two ways: (1) on individual scope tokens which is the standard evaluation and (2) on exact scopes in order to assess more strictly how efficient our models are for the labeling of all tokens correctly in each negation instance. For the latter, we use the error analysis script available online from previous work[i].

6. The cue detection task

Cue detection is the first step of the negation detection task. To tackle this problem, we experiment with two supervised learning approaches. First, a CRF model is trained using several features: words, lemmas, and part of speech tags, with a window over features which is defined empirically for each corpus. Our second approach uses a bidirectional LSTM with a CRF output layer, which is trained on the same features. We use embedding layers of dimension $k = 100$ with randomly initialized vectors for all features.

Table 4 presents the results obtained with our approaches on all four corpora. In all cases, the BiLSTM-CRF performs better than the CRF alone, which indicates that even on the task that appears to be simple enough for (non neural) machine learning methods, deep-learning methods can further improve the results. Indeed, the F_1 score obtained increases by up to three points.

Overall, the cue detection results are very high. However, our systems' results drop dramatically on Brazilian clinical trials. While cues such as *Falta de* were not found as no examples were available in the training set, the main reason for this drop lies in the lack of adjudication process. Indeed, as the discrepancies between annotators were not dealt with, they affect both precision and recall. For instance, a few occurrences of *anti-* were annotated as cues, but most were not, which caused recall errors. Moreover, precision errors were mostly forgotten cues such as *não*, *nehum* or *sem*.

7. The scope detection task

In all of the proposed scope detection experiments, the neural networks are trained only on sentences with negations. The base system takes an instance $I(n, c, t)$ as its input, where each word is

[i] https://github.com/ffancellu/NegNN

represented by: *n* vector with *word-embedding*, *c* vector with *cue-embedding*, *t* vector with *postag-embedding*. Cue and PoS-tag vectors are randomly initialized. We use embedding layers of 100 dimensions. For each system, we use the same empirically defined hyperparameters given before. During training, the embeddings weights are updated. Each of our corpora has been randomly segmented into the training set (80%, 20% for validation) and the test set (20%). We train and evaluate our systems on the *SEM-2012 datasets to get comparable results in English as well.

7.1 Results

Our first experiment compares the efficiency of both output layers (*softmax* and *CRF*) and three vector representations of words (random initialization, *word2vec* and *fastText*). Table 5 shows that using pre-trained word embeddings improves F_1 scores in most cases. On three out of four corpora, F_1 scores on scope tokens are similar for all output layer/word embeddings combinations (two points gap maximum). However, on Brazilian clinical trial protocols, the *softmax*/*fastText* combination gets significantly better results (78.20 while second best score is 74.87). Regarding the exact scope detection, the *CRF* output layer gets the best results on three out of four corpora. We expected such results because the *CRFs* are particularly efficient for tagging sequences. This hypothesis is confirmed by the results shown in Table 5, which position our results by comparison with the results obtained by other researchers on the *SEM-2012 data in English. Indeed, when computed on the **SEM-2012* data, our BiLSTM-CRF trained with FT-D gets a much higher F_1 than Li and Lu (2018) in terms of correctly labeled tokens (+1.27 points); however, we only get slightly higher results for the exact scope match (+0.41 points). Besides, in the medical domain, the returned scopes need to be as precise as possible and not to include medical concepts, which is more correctly managed by the *CRF* output layer. In both languages, we get better results on clinical data (clinical narratives and clinical cases) than on clinical trials. Overall, we get the best results on the CAS corpus with French clinical cases, with outstanding F_1 scores for the exact scope match. These results indicate that negative instances in clinical documents have simpler and more stable structures: typically, they contain less discontinuous cues and gaps in scopes. Moreover, we experiment with both recurrent neural network cells. As expected, the results in Table 5 indicate that the LSTM cells perform better than the GRU cells. Indeed, the review of the performances of RNNs on multiple tasks (Jozefowicz, Zaremba and Sutskever 2015) indicates that GRUs always show better results than LSTMs, except for the language modeling task. Although all gates have positive impact on the results, the experiments proposed in our work show that the forget gate gives the advantage to the LSTM. Regarding the exact scope detection, the BiGRU model benefits from the CRF ouput layer as well.

In our second set of experiments, in order to assess the efficiency of models trained on different sources of documents in a given language, we train the BiLSTM-CRF on one corpus (for instance clinical trials) and use another corpus (for instance clinical cases) to test this model. The results, shown in Table 5, indicate that clinical trial protocols offer more diverse instances of negation than clinical narratives. Indeed, the models acquired on clinical narratives perform poorly when compared to clinical trials. However, when trained on clinical trials, the models provide decent results for scope token detection, both in French and in Brazilian Portuguese. Regarding exact scopes, the results suffer from huge drops in all cases except when training on Brazilian clinical trials. Indeed, the best F_1 score on exact scopes from our first experiment on Brazilian protocols was 54.24, while now, when tested on clinical narratives, we get up to 73.30. Therefore, the poor results previously obtained on Brazilian protocols may be related to the fact that the testing set contains several instances of negation missing in the training set.

Figure 8 shows that 300–350 examples on average are necessary to achieve relatively good results. Adding more examples only slightly improves the results. Indeed, on French clinical trial protocols and Brazilian clinical narratives, using 650 training instances only improves the results

Table 5. Results for the scope detection task. The results are given in terms of Precision, Recall and F_1-score. The best scores are in bold

			Scope tokens			Exact scope match		
Corpus	System	WE[a]	P	R	F_1	P	R	F_1
ESSAI (French)	BiLSTM-S	RI	86.21	82.85	**84.50**	100	55.61	71.47
		W2V	83.54	83.68	83.61	100	56.59	72.27
		FT	80.79	86.41	83.51	100	56.59	72.27
	BiLSTM-CRF	RI	84.65	84.09	84.37	100	59.51	74.62
		W2V	83.86	83.10	83.48	100	61.95	**76.51**
		FT	82.38	84.84	83.59	100	59.51	74.61
	BiGRU-S	RI	81.52	86.25	83.82	100	52.68	69.01
	BiGRU-CRF	RI	83.12	84.42	83.76	100	57.07	72.67
CAS (French)	BiLSTM-S	RI	93.72	87.30	90.40	100	73.21	84.54
		W2V	93.03	88.69	**90.81**	100	75.59	86.10
		FT	91.50	88.69	90.08	100	72.02	83.74
	BiLSTM-CRF	RI	91.87	88.59	90.20	100	68.45	81.27
		W2V	91.47	88.29	89.85	100	76.19	86.49
		FT	94.82	87.10	90.80	100	78.57	**88.00**
Brazilian protocols (BRP)	BiLSTM-S	RI	74.40	70.77	72.54	100	30.23	46.43
		W2V	73.10	75.66	74.35	100	37.21	**54.24**
		FT	75.82	80.74	**78.20**	100	36.43	53.41
	BiLSTM-CRF	RI	78.02	67.25	72.24	100	24.81	39.75
		W2V	73.97	75.80	74.87	100	31.01	47.34
		FT	68.47	73.76	71.02	100	29.46	45.51
Brazilian narratives (BRN)	BiLSTM-S	RI	83.50	83.15	83.32	98.76	68.19	80.68
		W2V	82.46	81.88	82.17	98.73	66.76	79.66
		FT	83.67	81.56	82.60	99.59	69.80	82.08
	BiLSTM-CRF	RI	83.07	81.32	82.19	98.75	67.91	80.48
		W2V	85.97	81.17	83.50	98.82	71.92	**83.25**
		FT	88.72	81.17	**84.78**	98.82	71.92	**83.25**
*SEM-2012 (English)	Read *et al.* (2012)		81.99	88.81	85.26	87.43	61.45	72.17
	Lapponi *et al.* (2012)		86.03	81.55	83.73	85.71	62.65	72.39
	Packard *et al.* (2014)		86.10	**90.40**	88.20	98.80	65.50	78.70
	Fancellu *et al.* (2016)	W2V	92.62	85.13	88.72	99.40	63.87	77.70
	Li and Lu (2018) Semi *o*		94.00	85.30	89.40	100	69.10	81.70
	Li and Lu (2018) *Latent io*		94.80	83.20	88.60	100	69.50	82.00

Table 5. Continued

			Scope tokens			Exact scope match		
	BiLSTM+CRF	RI	**95.19**	84.40	89.47	99.46	69.58	81.88
	BiLSTM+CRF	FT-D	94.38	87.25	**90.67**	99.46	**70.34**	**82.41**
	BiLSTM+CRF	FT-O	94.37	84.40	89.11	99.41	64.26	78.06
Train–Test[b]	System	WE[a]	P	R	F_1	P	R	F_1
ESSAI–CAS	BiLSTM-CRF	RI	76.73	76.36	**76.54**	100	36.08	**53.03**
CAS–ESSAI		RI	82.36	55.23	66.12	100	28.20	43.99
BRP–BRN		RI	70.93	75.22	**73.01**	98.54	58.35	**73.30**
BRN–BRP		RI	77.37	49.53	60.40	100	25.35	40.45

[a] Word embedding: either random initialization (RI), *word2vec* (W2V) or *fastText* (FT) for French and Brazilian Portuguese; either random initialization (RI), a fastText model trained on Conan Doyle's novels (FT-D) or the original fastText model (1 million word vectors, FT-O) for English.
[b] Train: the corpus the system was trained on; Test: the corpus the system was tested on.

Figure 8. Learning curve for the BiLSTM-CRF, without pre-trained Word Embeddings.

by approximately 1.5 points. This Figure also indicates that negation detection can be further improved on Brazilian clinical trial protocols when more reference data are available.

7.2 Error analysis

Although our results on most corpora seem consistent, our results drop dramatically on Brazilian clinical trials. Once again, as conflicts between annotators were not resolved in this case, many tokens end up incorrectly annotated as being part of a scope. Indeed, in several cases, tokens surrounding *anti-* were annotated as scope tokens, even with cue embeddings signaling the absence of *anti-* as a cue.

In order to study the frequent types of errors, a large portion of sentences containing at least one prediction error was manually examined and the causes of error were annotated. In the examples below, the negation cues are underlined, the scopes are in bold, and the predictions are between

brackets. In Example 1, the prediction fails at labeling ***rénale****(renal)*. This problem of missed adjectival attachment is due to the fact that, in the majority of cases in the reference data, the scopes associated to the cue ***sans*** often only include one token, which may be causing this error that impacts recall:

1. *Le patient sortira du service de réanimation guéri et sans* ***[insuffisance] rénale*** *après huit jours de prise en charge et cinq séances d'hémodialyse.*
   (The patient will be discharged from the intensive care unit without renal failure after eight days of management and five hemodialysis sessions.)

This type of error also occurs with prepositional attachment that are missed (impacting recall) or wrongly included in the scope (impacting precision). It occurs especially with the French preposition *de* and the Brazilian one *da* as in the following examples in which *da última consulta* was wrongly predicted as part of the scope:

2. *NEGA [****INTERCORRÊNCIAS*** *DA ÚLTIMA CONSULTA] PRA CÁ*
   (Denies intercurrent disease from the last consultation to now)

Example 2 illustrates the error that impacts precision. Here, the model wrongly predicts that all tokens in the sentence are within the scope. In the reference data, the cue ***aucun*** *(any, no)* often occurs at the beginning of sentences and in sentences with many instances of negation. The model, mostly trained on this kind of examples, may try to reproduce these structures which cause bad prediction in some cases.

3. *[Les colorations spéciales (PAS, coloration de Ziehl-Neelsen, coloration de Grocott]) ne* ***[mettaient en évidence]*** *aucun* ***[agent pathogène]****.*
   (Special stains (PAS, Ziehl-Neelsen stain, Grocott stain) showed no pathogens.)

In Example 3, the error impacts both precision and recall. In this example, we have two instances of negation with the same cues: ***n****. . .****pas****.* Usually, its scope follows, however, in the first instance it precedes. As we do not have many examples of this kind to train on, the model fails to correctly label the sequence. In the second negation instance, the scope may be shorter than usual, which impacts precision.

4. ***Le retrait du matériel d'ostéosynthèse incriminé*** *n'[****est****] pas [****systématique****], ce qui explique qu'il n'[****ait****] pas [****été proposé*** *à notre patient asymptomatique].*
   (The removal of the implicated osteosynthesis material is not systematic, which explains why it has not been proposed to our asymptomatic patient.)

The error illustrated by Example 4 indicates that the cues *nul* and *a* are underrepresented in our corpus. Indeed they only occur once. Therefore, this model was not trained on them and thus assigns a partially incorrect scope.

5. *A.* ***Tracé*** *nul [****et****] a[****réactif****].*
   (Null and unresponsive route)

The errors illustrated by the next examples are due to errors in annotation. Indeed, performing an error analysis is useful to detect the annotation mistakes as well. Thus, in the following example, the model predicts every token correctly. However, the comma was wrongly annotated as part of the scope. Overall, this error affects both recall and exact scope match scores.

6. *La patiente ne* ***[fume]*** *pas, ne prend que très rarement de l'alcool et n'[****a****] pas* ***[d'allergie aux médicaments]****.*
   (The patient does not smoke, only rarely takes alcohol and has no drug allergy.)

The last example on French presents the annotation error as well. In this case, ***ce*** *(this)* and ***est*** *(is)* should not be within the scope since ***n'*** is not a cue.

7. ***Ce*** *n'****est*** *que 48 heures après la dernière dose que les troubles visuels et les hallucinations disparaissent complètement, sans [****laisser de séquelles****].*
(It is only 48 hours after the last dose that the visual disturbances and hallucinations disappear completely, without leaving any sequelae.)

The types of errors found in Brazilian corpora prove to be even more complex. Indeed, we find that a large number of predictions combine precision and recall errors within a single sentence. The examples below illustrate this situation:

8. Ausencia de diagnóstico [de **doenças neuromusculares**], [**trauma**], **tumores** ou [**abscessos raquimedulares**], **hemiplegia/ paresia**, **lesão de plexo** ou [**encefalopatia**] **cerebral**.
(Absence of diagnosis of neuromuscular diseases, trauma, spinal tumors or abscesses, hemiplegia/paresis, plexus injury or cerebral encephalopathy.)
9. que não apresentem [outras **doenças neurológicas**] ou [**ortopédicas**] **diagnosticadas**.

(Does not have other diagnosed neurological or orthopedic diseases.)

The error in Example 8 corresponds to the inclusion or not of function words (*de, outras*) within the scope of the markers. Another error (Ex. 9) is related to the completeness of nominal groups (*encefalopatia cerebral* and *ortopédicas diagnosticadas*).

Last, for English, the errors are also due to similar problems (eg. prepositional attachment). Compared to other approaches, our system solves errors committed by Fancellu *et al.* (2016), such as Example 10 in which the scope predicted by their system includes the main predicate with its subject in the scope.

10. You felt so strongly about it that [I knew **you could**] not [**think of Beecher without thinking of that also**].

Moreover, by running their code, we can find several errors committed by their system but correctly resolved by our system, such as in the following examples. However, the predictions we get from their system may not be identical to their best predictions since the word2vec model they use is unavailable.

11. [Just the word], nothing [**more**].
12. "Well, **can** [**you give**] **me** no [**further indications**] ?"

8. Conclusion and future work

The interest for automatic detection of negation in English with supervised machine learning has increased in the recent years. Yet, the lack of data for other languages and for specialized domains hampers the further development of such approaches. In this work, after presenting the difficulties related to this task and after a brief reminder of existing work, we presented new bodies of biomedical data in French and Brazilian Portuguese annotated with information on the negation (cues and their scope). Prior to the dissemination to the research community, the French and Brazilian clinical trial protocols corpora will be finalized through the integration of new data and the computation of the IAA. The French CAS corpus will be distributed as more automatically annotated sentences are corrected. The Brazilian corpus with clinical narratives may prove to be more difficult to share with the community, as it contains data on real patients.

Another contribution of our work is the exploitation of different types of word vector representations and recurrent neural networks for the automatic recognition of the cues and scope

of negation. The experiments are conducted and evaluated in three languages: in English, which has independent reference data, and in French and Brazilian Portuguese, which have not experienced much work of this type, especially for the biomedical domain which contains specific negation phenomena. Our system yields state of the art results on **SEM-2012*, which validates our approach. It also shows good performances on our corpora except for the Brazilian clinical protocols. From a more technical point of view, our work also indicates that the LSTM-based neural architectures are more efficient than GRUs in the scope detection task, although the latter are now preferred in many other NLP tasks. In addition, the *CRF* layer brings better performance than the *softmax* on exact scope match which is of importance to us given the need to correctly assess which medical concepts are present or absent. Finally, the models are applied to general-language and medical data.

In the future, we plan to extend the system to the detection of uncertainties and of their scope, whose setting is very similar to the negation task. Besides, we plan to improve our neural network performance by providing richer feature set. In particular, recent embedding techniques, such as BERT or ELMo (Devlin *et al.* 2018; Peters *et al.* 2018), may provide more accurate representation of the sentences. Syntactic parsing of sentences may also provide useful features for scope detection.

Financial support. This work was partly funded by the French government support granted to the CominLabs LabEx managed by the ANR in Investing for the Future program under reference ANR-10-LABX-07- 01.

References

Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mane D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viegas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y. and Zheng X. (2016) TensorFlow: Large-scale machine learning on heterogeneous distributed systems. pp. 1–19.

Abdaoui A., Tchechmedjiev A., Digan W., Bringay S. and Jonquet C. (2017) French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français. 4e édition du Symposium sur l'Ingénierie de l'Information Médicale. Toulouse, France, pp. 1–10.

Albright D., Lanfranchi A., Fredriksen A., Styler IV W.F., Warner C., Hwang J.D., Choi J.D., Dligach D., Nielsen R.D., Martin J., Ward W., Palmer M. and Savova G. K. (2013) Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association* **20**(5), 922-930.

Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Bojanowski P., Grave E., Joulin A. and Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Chapman W., Bridewell W., Hanbury P.F. Cooper G. and Buchanan B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* **34**, 301–310.

Cho K., van Merriënboer B., Gulcehre C., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734.

Deléger L. and Grouin C. (2012). Detecting negation of medical problems in French clinical notes. In *IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA, pp. 697–702.

Denny J.C. and Peterson J.F. (2007). Identifying QT prolongation from ECG impressions using natural language processing and negation detection. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. Studies in Health Technology and Informatics, Vol. 129, IOS Press, pp. 1283–1288.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**(6), 391–407.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019,* Minneapolis, Minnesota, June 2–7, 2019. Association for Computational Linguistics, pp. 4171–4186.

Elkin P.L., Brown S.H., Bauer B.A., Husser C.S., Carruth W., Bergstrom L.R. and Wahner-Roedler D.L. (2005). A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making* **5**, 13.

Fancellu F., Lopez A. and Webber B. (2016). Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Volume 1, Berlin, Germany: Association for Computational Linguistics, pp. 495–504.

Gindl S., Kaiser K. and Miksch S. (2008). Syntactical negation detection in clinical practice guidelines. *Studies in Health Technology and Informatics* **136**, 187–192.

Grabar N., Claveau V. and Dalloux C. (2018, October). CAS: French corpus with clinical cases. In *LOUHI 2018: The Ninth International Workshop on Health Text Mining and Information Analysis*. Brussels, Belgium: Association for Computational Linguistics, pp. 122–128.

Harkema H., Dowling J.N., Thornblade T. and Chapman W.W. (2009). ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* **42**(5), 839–851.

Hartmann N., Fonseca E., Shulby C., Treviso M., Rodrigues J. and Aluisio S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Proceedings of Symposium in Information and Human Language Technology*, Uberlândia, MG, Brazil, October 2–5, 2017. Sociedade Brasileira de computação.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural computation* **9**(8), 1735–1780.

Jozefowicz R., Zaremba W. and Sutskever I. (2015, June). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pp. 2342–2350.

Lafferty J., McCallum A. and Pereira F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282–289, San Francisco, CA, USA.

Lapponi E., Velldal E., Øvrelid L. and Read J. (2012, June). Uio 2: Sequence-labeling negation using dependency features. In *Proceedings of 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*. Montréal, Canada: Association for Computational Linguistics, pp. 319–327.

Li H. and Lu W. (2018, July). Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics, Volume 2, pp. 533–539.

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*. Lake Tahoe, Nevada: Curran Associates Inc. Red Hook, NY, USA, pp. 3111–3119.

Morante R., Schrauwen S. and Daelemans W. (2011). Annotation of Negation Cues and their Scope. Guidelines v1.0. Computational linguistics and psycholinguistics technical report series, CTRS-003, pp. 1–42.

Mutalik P.G., Deshpande A. and Nadkarni P.M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *Journal of the American Medical Informatics Association* **8**(6), 598–609.

Névéol A., Grouin C., Leixa J., Rosset S. and Zweigenbaum P. (2014). The Quaero French medical corpus: A ressource for medical entity recognition and normalization. In *Proc BioText M*, Reykjavik, Iceland: Citeseer.

Nguyen D.Q., Nguyen D.Q., Pham D.D. and Pham S.B. (2014). RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 17–20.

Oliveira L.E.S., Peters A.C., Da Silva A.M.P., Gebeluca C.P., Gumiel Y.B., Cintho L.M.M., Carvalho D.R., Hasan S.A. and Moro C.M.C. (2020). SemClinBr – a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks. arXiv preprint.

Packard W., Bender E.M., Read J., Oepen S. and Dridan R. (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, pp. 69–78.

Peng Y., Wang X., Lu L., Bagheri M., Summers R. and Lu Z. (2018). NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings* **2017**, 188–196.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018,* pp. 2227-2237. New Orleans, USA, June 1–6, 2018.

Read J., Velldal E., Øvrelid L. and Oepen S. (2012, June). Uio 1: Constituent-based discriminative ranking for negation resolution. In *Proceedings of 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*. Montréal, Canada: Association for Computational Linguistics, pp. 310–318.

Rehurek R. and Sojka P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. Citeseer.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing,* pp. 44–49. Manchester, UK.

Uzuner Ö., South B.R., Shen S. and DuVall S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5), 552–556.

Velldal E., Øvrelid L., Read J. and Oepen S. (2012). Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics* **38**(2), 369–410.

Velupillai S., Dalianis H. and Kvist M. (2011). Factuality levels of diagnoses in Swedish clinical text. In *Proceedings of the Medical Informatics Europe conference 2011 - The XXIIIrd International Congress of the European Federation for Medical Informatics.* Oslo, Norway: Studies in Health Technology and Informatics 169, IOS Press, pp. 559–563.

Vincze V., Szarvas G., Farkas R., Móra G. and Csirik J. (2008). The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* **9**(11), S9.

Cite this article: Dalloux C, Claveau V, Grabar N, Oliveira LES, Moro CMC, Gumiel YB and Carvalho DR. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering* https://doi.org/10.1017/S1351324920000352