View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Archive

THE UNIVERSITY of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Computational models for multilingual negation scope detection

Federico Fancellu

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2018

Abstract

Negation is a common property of languages, in that there are few languages, if any, that lack means to revert the truth-value of a statement.

A challenge to cross-lingual studies of negation lies in the fact that languages encode and use it in different ways. Although this variation has been extensively researched in linguistics, little has been done in automated language processing. In particular, we lack computational models of processing negation that can be generalized across language. We even lack knowledge of what the development of such models would require.

These models however exist and can be built by means of existing cross-lingual resources, even when annotated data for a language other than English is not available. This thesis shows this in the context of detecting string-level negation scope, i.e. the set of tokens in a sentence whose meaning is affected by a negation marker (e.g. 'not'). Our contribution has two parts.

First, we investigate the scenario where annotated training data is available.

We show that Bi-directional Long Short Term Memory (BiLSTM) networks are state-of-the-art models whose features can be generalized across language. We also show that these models suffer from genre effects and that for most of the corpora we have experimented with, high performance is simply an artifact of the annotation styles, where negation scope is often a span of text delimited by punctuation.

Second, we investigate the scenario where annotated data is available in only one language, experimenting with model transfer.

To test our approach, we first build NEGPAR, a parallel corpus annotated for negation, where pre-existing annotations on English sentences have been edited and extended to Chinese translations.

We then show that transferring a model for negation scope detection across languages is possible by means of structured neural models where negation scope is detected on top of a cross-linguistically consistent representation, Universal Dependencies. On the other hand, we found cross-lingual lexical information only to help very little with performance. Finally, error analysis shows that performance is better when a negation marker is in the same dependency substructure as its scope and that some of the phenomena related to negation scope requiring lexical knowledge are still not captured correctly.

In the conclusions, we tie up the contributions of this thesis and we point future work towards representing negation scope across languages at the level of logical form as well.

Acknowledgements

First and foremost, I'd like to thank both my supervisors, Bonnie Webber and Adam Lopez. I cannot even begin to imagine what this thesis, or my entire Ph.D. experience would have been like if it wasn't for their constant support and guidance.

I also want to thank Prof. Lee Jiyoung, who supervised my first Master's thesis for teaching me what good research should look like.

I would also like to thank people whom I collaborated with for parts of this work. Qianchu Liu who collaborated with me in annotating NegPar was the source of a lot of interesting discussion on Chinese. With Hangfeng He, we worked on negation detection in Chinese and were able to better understand the model we experimented with. As for Siva Reddy, I will always cherish the stimulating discussions we had.

All the people of the Sassy Soc. deserve a special mention. Ph.D. wouldn't have been the same without our trips to the Picnic Basket or Friday's cookie time on a reduced budget. Thank you Clara, Craig, Ida, Joana, Marco, Naomi, Nik, Sameer, Sorcha and many more. A special thanks goes to Mihai who I have shared a lot of Ph.D. related frustrations with.

Yet another special mention goes to the fine gentlemen I had the chance to work with during the Alexa challenge; Joachim, Daniel, Ben and Manny (along with Marco and Mihai), I have really learnt a lot from you.

Finally on a more personal note, I'd like to thank my parents for always being there for me. Much of whom I am today is because of them.

The last and the most important acknowledgment goes to my partner, Adam. Nothing would have been possible without you (not even formulating this sentence correctly with two instances of negation...). Your support and care throughout this experience was the best remedy to my gloomy days. And sorry for my tantrums, you handled them in the best way possible.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Federico Fancellu*)

Table of Contents

Chapter 1

Introduction

Negation is a fundamental property of human language, enabling speakers to reverse the truth value of a statement. Negation is also a common property of languages, in that there are few, if any, languages that lack negation. For this reason, perhaps unsurprisingly, automatically processing negation is a task that has been widely investigated in the NLP community.

This thesis focuses on one key component of negation, negation scope, i.e. the set of elements falling under the negation operator. Whereas most semantic banks represent negation in an abstract meaning representation that is First-Order Logic (FOL)-equivalent or translatable (e.g. DeepBank (Flickinger et al., 2012), Groeningen Meaning Bank (Bos et al., 2017a)), where the scope is the set of predicates falling under the negation operator, another research thread has focused on anchoring the semantics of negation to strings. The dichotomy between logical-form negation and string-level negation is exemplified in (1), with the negation marker in **bold** and the scope underlined.

(1) Cats do not like pizza

1. $\neg\exists x.e.y.cat(x) \wedge like(e) \wedge pizza(y) \wedge arg0(e,x) \wedge arg1(e,y)$
2. Cats do **not** like pizza

In (1) the scope spans the entire formula; the string representation mirrors the scope of the logical form in that the operator is anchored to a negation cue (‘not’) and we mark ‘cats’, ‘do’, ‘like’ and ‘pizza’ as being part of the negation scope.

In this thesis we focus on string-level negation, investigating in depth the task of **detecting negation scope from raw text**. We do this because there are certain advantages to be gained by operating at a string-level. Anchoring a semantic phenomenon to strings can ease its processing, which is why negation scope detection has found numerous applications including biomedical information extraction (Morante et al., 2008) and machine translation (Fancellu and Webber, 2014). We will move instead towards a formal semantic representation of negation scope when discussing future directions.

However, previous work has been limited to focusing mostly on English or on the development of language-dependent models with little or no exploration of how languages vary in representing negation and what methods could help with detecting it.

This thesis fills this gap by looking at negation scope from a multilingual perspective. We emphasize that detecting (as well as representing) negation scope automatically in different languages raises related *data* and *modeling* problems. The data problem lies in the fact that there exists only one corpus annotated for negation in a language other than English and that annotation guidelines are not consistent across corpora. The modeling problems refers to the lack of computational models to process negation that can be generalized across languages, as well as to the lack of knowledge of what the development of such models requires.

We narrow down these problems to two main research questions:

1. If annotated data for multiple languages is available, is there a model that is general enough to be applied to different languages?

2. If annotated data is available only in one language, could the annotations be projected onto a multi-lingual representation so that, by training the model in a language, it can be transferred onto another?

To answer both questions, we focus on two languages, English and Chinese, which are the only languages for which data annotated for negation scope already exist. This thesis is structured as follows:

We introduce relevant background in Chapter 2.

Chapters 3-5 constitute the main body of this thesis. To describe the contribution of each chapter as well as to summarize their content, we break down the two research questions highlighted above into smaller ones.

Question 1 is the focus of Chapter 3.

Can we build a scope detection model relying only on features that can be generalized across languages? Yes; we found Bi-directional Long-Short Term Memory (BiLSTM) networks to be generalizable models yielding state-of-the-art performance while relying only on two embedding features, word and universal PoS tags. This contrasts with previously developed classifiers, which are highly-engineered and English-specific, relying on features and hand-crafted heuristics that are not applicable to other languages.

How robust are recurrent neural architectures across different corpora in English and Chinese? In both languages, BiLSTMs yield state-of-the-art performance. However, the performance of these systems suffer when train and test sets are of different genres, often failing to outperform a non-neural classifier using syntactic features. We also found that syntax affects the performance of negation scope detection, with VP negation being overall easier to detect than other kinds.

When training and testing on the same corpus, we report two main findings.

First, when negation scope is a continuous sentence span, we found predictions to leave gaps in between. To tackle this issue, we found that adding a transition based component on top of the BiLSTM model ensures prediction to be continuous as well.

Second, we found that high performance is often due to negation scope being annotated as a single span of text delimited by punctuation. For negation scopes not of this form, detection accuracy is low and under-sampling the easy training examples does not substantially improve accuracy. We demonstrate that this is partly an artifact of the annotation style, and we argue that future negation scope annotation efforts should focus on these more difficult cases.

Chapter 4 and 5 will focus on Question 2.

Can we build a model trained on English that is directly transferable across languages? First, we need means to evaluate this method that could also be compared to an oracle monolingual systems. This means that we require data annotated in the same style for both English and a target language, whereas up-to-now no available corpus annotated for negation scope meets this requirement.

To remedy this, we develop NEGPAR, a parallel English-to-Chinese corpus annotated for negation. The corpus was created by leveraging pre-existing English annotations (CONANDOYLENEG, Morante and Daelemans, 2012) which are projected to a Chinese translation and then manually corrected. We show that the annotations in the CONANDOYLENEG corpus fall short with respect to the phenomena they consider, reason why we edit these prior to projection. At the same time, we also develop annotation guidelines for Chinese that we hope will encourage future work. Finally, we show that annotation projection via word-alignment does not help substantially, due to both differences in translation and alignment errors.

Using NEGPAR, we then attempt to train a model in a source language and use it in a target one where data annotated for negation scope might not be available. To bridge the gap between these two languages we use two intermediate representations: Universal Dependencies (De Marneffe et al., 2014), a syntactic annotation framework consistent across languages, as well as cross-lingual word embeddings.

We experimented with structured neural classifiers, namely a bi-directional Dependency LSTM (modeled after the treeLSTM of Tai et al. (2015)) and Graph Convolutional Networks (Marcheggiani and Titov, 2017), that are able to classify negation scope on top of UD trees while ignoring word order and other linear information. We compare their performance to our state-of-the-art BiLSTM.

Results show that it is indeed possible to build a cross-lingual model for negation scope detection, although its performance is not as good as a monolingual oracle. We found that structure is what matters the most for the task of negation scope detection, whereas word embeddings information does not improve or on the contrary, hinders performance. This is particularly evident in the performance of structured classifiers that are able to leverage information from the UD tree alone to predict the tokens in the scope. When this structural information is absent as in the case of the BiLSTM, the networks uses once again punctuation and sentence boundaries to guide prediction.

Through an error analysis we also show that structured classifiers are better in predicting scope that is in the same dependency substructure as the negation marker and that they fail to predict some of the phenomena related to negation scope where lexical information is needed.

Chapter 6 contains the conclusions of this thesis, where we tie up the aforementioned contributions in relation to the overarching goals of this thesis. Future research directions are discussed relating to Question 2. but considering this time the formal semantics of negation scope.

Summary of the contributions. We summarize the findings of this thesis as follows.

- For the task of *monolingual* negation scope detection, BiLSTMs are state-of-the-art models that can be generalized across language.

- In most corpora annotated for negation scope, punctuation boundaries are already a strong baseline for negation scope detection. Punctuation boundaries also guide prediction in the BiLSTM model.

- Adding a transition-based component on top of a BiLSTM helps in predicting continuous scopes.

- In the presence of English sentences annotated for negation scope and their translations in a target language, projection via word alignment information has a low recall, showing that most negation instance in Chinese are rendered in English as positive construction or are lost in the alignment noise.

- Negation scope can be detected across languages on top of Universal Dependencies parses, since syntactic annotations are not language-specific.

- Structured neural models are the best fit for the task of cross-lingual negation scope detection, with bi-directional recursive LSTM outperforming Graph Convolutional Networks; on the other hand, BiLSTM overfit and once again rely mostly on punctuation to guide prediction.

- Cross-lingual word embeddings only contribute little to the performance of a cross-lingual model.

- Structural models perform better when negation scope are in the same dependency subtree as the negation marker and still fail in handling some phenomena related to negation scope correctly, especially when lexical information is required.

- The findings of this thesis can be used to inform future work on representing negation scope across languages in a logical form as well.

We have also produced the following two resources:

- NEGPAR: a parallel English-to-Chinese corpus annotated for negation, alongside annotation guidelines for Chinese.

- NEURALNEG: a suite of neural network models for negation scope detection at the string level

Chapter 2

Background

2.1 Why negation?

Negation is a universal linguistic device, in that every language possesses a way to reverse the truth value of a statement or parts of it. Negation can be used to negate entire sentences (‘She does not eat pizza’), to bring the focus on particular portions of it (‘I am coming but not tomorrow’) or it can be used pragmatically (’I don’t like it – I love it.’). Furthermore the presence of negation affects how we understand a statement; for instance, as shown in (2), negating the main predicate triggers an inversion in the entailment direction.

(2) 1. I am eating meat
$\not\models$ I am eating red meat
2. I am not eating meat
$\models$ I am not eating red meat

Negation is a well-understood phenomena in linguistics, i.e. the way we express negation and how it relates to other meaning components are well-documented. From works as old as Aristotele’s *Categories* and *De Interpretatione*, negation have been the focus of innumerous studies (Horn, 1989; Pullum and Huddleston, 2002; Ladusaw, 1979, amongst many) that have looked at the form and the meaning of negative statements as well as the interaction with other operators, such as universal quantification and modality. Studies have also discussed its interaction with the surrounding syntactic context (e.g. the difference between *sentence* and *constituent* negation raised by Klima (1964) and Jackendoff (1969), which we will discuss more in depth in later sections) as well as its interpretation with respect to the entities mentioned in a proposition (e.g.

whether one should consider 'the king of France' in the scope of negation in the sentence "The king of France isn't bald' where there isn't any king of France" (Russell, 1905)).

However, the interest in automatically processing negation is fairly recent.

On one hand, representing negation and its scope in a FOL-translatable logical form, as it is the case of different semantic banks, e.g. the Groeningen Meaning Bank (Bos et al., 2017b) and DeepBank (Flickinger et al., 2012), allows for a large-scale analysis and processing of its interaction with other semantic phenomena.

However, it is grounding and processing negation at a string level that has increased the interest in automatically detecting negation from text in a variety of NLP domains.

In the biomedical domain, detecting negation was found to be of great importance when automatically processing medical records, including patient medical histories and radiology reports. Whereas the first systems for negation detection were rule-based, e.g. NegEx (Chapman et al., 2001) and Negfinder (Mutalik et al., 2001), with patterns tailored to fit negation in the context of medical records, it is only with the annotation of the BIOSCOPE corpus (Vincze et al., 2008, 2011) that negation scope detection became mostly a supervised machine learning task.

The importance of processing negation is not restricted to the biomedical domain. In sentiment analysis, it is perhaps unsurprising that recognizing the span affected by negation helps with detecting the sentiment of a document (see Wiegand et al. (2010) for a survey of different approaches). However, this requires in-domain corpora: Konstantinova et al. (2012) present a corpus of product reviews, while Reitan et al. (2015) present a corpus of tweets annotated for negation.

Interest in negation in other or across languages has also started to emerge in recent years. To date, there exists only one corpus annotated for negation in a language other than English, the Chinese Negation and Speculation Corpus (CNeSp, Zou et al. (2016)). Nevertheless, work in Statistical Machine Translation has shown for several language pairs that translating negation is an issue and that one could benefit from a semantic representation of it. Negation poses a *modelling* challenge in that languages differ in the way they signal negation, whereas machine translation systems are often agnostic to semantics (Baker et al., 2012; Fancellu and Webber, 2014, 2015). Negation also poses a problem of *data*: humans make more positive statements then negative ones and corpora reflect this property where affirmative outnumber negative sentences (Wetzel and Bond, 2012).

2.2 How is negation expressed across languages?

It is widely accepted that every language has a way to reverse the truth value of a statement but languages differ greatly in how they do it. The purpose of this section is to assess whether it is possible to restrict this variation to a finite set of patterns and to this end, we present a survey of language typology studies on this topic. In particular, we will refer mainly to the work of Payne (1985), Dryer (2005) and Miestamo (2007) on the topic.

A first distinction is made between negation on the predicate of a main sentence (either verbal or adjectival) and negation expressed elsewhere. More formally we can distinguish between *sentential* (or "standard") and *non-sentential* negation: the former refers to those instances of negation applicable to negation on the main predicate (either in verbal or in copular constructions) whereas the latter to negation expressed on adverbs, quantifiers or elements other than the predicate. For instance in 'Students do **not** eat pizza' negation is sentential, whereas in '**Not** every student eats pizza', it constrains the universal quantifier. Another axis of variation relates to the way a language expresses negation in declarative verbal main sentences as opposed to other negative clausal environments such as imperatives, existentials and non-verbal negation.

When describing the realization of standard negation, most works seem to agree on the main differentiation between *morphological* and *syntactic* negation, which was also the focus of earlier work, such as Dahl (1979).

Morphological negation can be divided into sub-categories according to the position of the negation-bearing morpheme with respect to the verb: prefixal (as shown for Czech in (3)), suffixal or circumfixal.

(3) **Ne**rozumím
NEG.understand.PRES.1PERS.SING
'I do**n't** understand'

While Payne (1985) claims no language can encode negation by means of word order or intonation, Miestamo (2007) includes in this classification prosodic and reduplicative morphological negation as marginal, rarely attested categories.

Syntactic negation can be realized by means of a lexical element (e.g. English in (4.1)) or an auxiliary verb that may or may not be inflected (e.g. Finnish in (4.2) and Estonian in (4.3)).

(4) 1. English: He is **not** reading.

2. Finnish: **En** tiedä
Not.1PERS.SING Know.IND.PRES
'I do**n't** know'

3. Estonian: **Ei** loe
Not read.IND.PRES
'I/you/he/she/we/they do **not** read'

In the case of an inflected auxiliary verb, the features for which the auxiliary is inflected are usually not repeated on the main verb so as to avoid redundancy.

Finally, syntactic negation can be classified according to its position with respect to the negated verb; negation can in fact be pre-verbal, post-verbal and circumverbal (e.g. French and Welsh). Dryer (2005) also adds that in languages like Korean, both lexical and morphological standard negation coexist. Coexistence of lexical standard negation and constituent-based morphological negation can also be observed in languages like English, although one can argue that in English affixes like 'un-'('known') and 'im-'('patient') create contraries, not contradictions.

As for non-standard negation, certain languages realize it in a discontinuous manner where two parts of the same expression encapsulate the span of the sentence negated. In later chapters, we will see this is the case of the 'except' construction in Chinese, where two parts of the same expression '除了...以外' appear at the start and at the end of the negated clause.

(5) 除了 他 以外 ， 我们 都 是 学生
Except he except , we all be student
'We are all students, except him'

Another important aspect of typological studies is the contrast between positive and negative statements. A first classification between *recusative negation* (négation récusative) and *suspensive - reassertive negation* (négation suspensive-réassertive) is reported in Forest (1993), where the former indicates a negative statement identical to its positive counterpart except for the presence of the negation marker, while the latter describes those negative statements where one or more grammatical domains are marked differently than the positive (suspensive), while being in the declarative realm (reassertive). Similarly we can oppose *symmetric* and *asymmetric* negation. In symmetric negation instances, the sentence preserves the same elements of its positive counterpart with only the addition of the negation-bearing elements. Asymmetric negation implies on the other hand that certain morphological features are deleted or reduced when transforming a positive sentence into a negative one. This difference is exemplified by Dutch (ex. 6) and Korean (ex. 7).

(6) Ik zing
I sing.PRES.1PERS.SG
'I sing'
Ik zing **niet**
I sing.PRES.1PERS.SG not
'I do**n't** sing.'

(7) jeo saramui ireumeul arayo
that person.GEN name.ACC know
'I know the name of that person'
jeo saramui ireumeul mollayo
that person.GEN name.ACC not know
'I do**n't** know the name of that person'

In Dutch, when inverting the truth value of a sentence, all elements are preserved and the negation marker is inserted. On the other hand, in Korean there are verbs like 'alda', 'to know', that have a dedicated negative counterpart (in this case 'moreuda', 'to not know') and do not allow for a negation marker to precede the verb.

Miestamo (2007) also takes into consideration imperative, existential and non-verbal negation as examples of clausal environments where asymmetry is the most evident. In the case of imperatives, a four-way categorisation is proposed according to whether the prohibitive verbal construction follows the 2nd singular imperative and whether negation is expressed the same way as in declarative sentences. For instance, as we will see in later chapters, in English negation on commands is expressed the same way as declarative sentences with 'not' preceded by the 'do' support; on the other hand, Chinese has a dedicated marker to use only in imperative forms, 别, which contrasts with the one used in declarative sentences, 不. This contrast is exemplified below:

(8) 她 不 动
She not move
'She does**n't** move'
别 动
not.IMP move
'Do **not** move'

In the case of existential clauses on the other hand, negation can be expressed by means of (i.) a negative marker attached to a separate existential verb (ii.) a negative existential verb or (iii.) the standard negation marker, which takes on the role of negative existential. Finally in the case of non-verbal negation, Miestamo (2007) considers negation in nominal predicates, distinguishing those languages where negation is expressed on the copula dominating a nominal complement (e.g. English, Thai) and languages where negation is expressed directly on the nominal element. Payne (1985) reports other contrasts such as change of word order, tonal change, the insertion of supportive verbs (such as the English *do*) and change of words surrounding negation (e.g. Russian, where the accusative case for the direct object in certain cases becomes genitive).

As for non-standard negation, Payne (1985) distinguishes between:

- Negated quantifiers (e.g. English, *not many*): usually in pre-verbal position, although some languages also allow negated quantifiers to appear in post-verbal position.

- Inherently negated quantifiers (e.g. English, *anything*): where a distinction is made between languages that require negation with these quantifiers (giving rise to the double negation phenomenon that Dryer (2005) also mentions) and languages that do not require negation with these quantifiers. Regarding this category, Miestamo (2007) mentions Haspelmath (2005)'s three-way distinction based on whether the presence of these quantifiers (or indefinite negative pronouns, in his terms) requires negation on the predicate of the clause they belong to or its presence is optional.

- Negated adverbials (e.g. English, *not often)*: the syntactic position of this category varies between languages. In English, they are allowed only in pre-verbal position whilst in other languages, such as Russian, they can be post verbal.

- Inherently negative adverbs (e.g. English *never*): they are not as syntactically restricted as negated adverbials and can also appear post-verbally.

To conclude, in some languages the presence of negation has an effect on the surrounding context, 'attracting' certain words or 'inhibiting' the presence of others. *Negative concord* (e.g. Labov (1972); Baker (1970) amongst others) is the phenomenon whereby negative indefinite pronouns co-occur with sentential negation. As shown in the example below, the Italian pronoun *mai* (English ‘never’) requires sentential negation, whose absence would lead to an ungrammatical sentence.

(9) Io **non** vado mai al cinema (* Io vado mai al cinema)
I not go-PRES-1PERS-SING never to the cinema
‘I never go to the cinema’

English displays a similar negative concord in certain varieties of English; as shown in the following example of Alabama English, which allows ‘never’ to appear with the ‘not’ Feagin (1979).

(10) I ai**n’t** never been drunk (‘= I have never been drunk.’)

Similar to negative concord *negative polarity items(NPI)* (e.g. Klima (1964); Ladusaw (1980); Haspelmath (2001) amongst others) are lexical items licensed by sentential negation. In English these includes pronouns (e.g. ‘any’, ‘either’) or expressions (e.g. ‘lift a finger’) that can only co-occur with an instance of negation.

(11) I have**n’t** bought any ever since (*I have bought any ever since)
He hasn’t lifted a finger (?He has lifted a finger)

2.3 Negation in English and Chinese

Based on the theoretical insights of the previous section, how do English and Chinese differ in the way they represent negation?

In general, English and Chinese display the same syntactic order (Subject-Verb-Object) with both sentential and non-sentential negation expressed similarly. As shown in example (12), ‘not’ and ‘不’ are both negation adverbs placed before a VP to flag sentential negation, whereas the morphemes ‘im-’ and ‘不’ are affixed to the adjectives ‘patient’ and and ‘耐心‘ to negated an adjective inside an NP (13).

(12) 我 不 吃 猪肉
I not eat pork
‘I do**n’t** eat pork’

(13) 他 是 一位 不耐心 的 人
He is one-CL. impatient of man
‘He is an **im**patient man’

However, defining the case in (13) as morphological negation in Chinese is as problematic as defining what a word is, given that no word boundaries are present in written texts. ‘不耐心‘ can be in fact be considered as a whole with the marker being affixed to the adjective (as a matter of fact, ‘不耐心‘ figures as an entry in dictionaries) or as two separate words, since both the marker and the adjective can both stand as independent words. We will come back on the semantic status of morphological negation in Chinese in Ch. 4.

If we define these cases as morphological negation, then unlike English, the negation cue in Chinese can be infixed. This is the case in resultative constructions where a verb is followed by a complement indicating potentiality, directionality or result; if negated, as shown in (14), the cue appears infixed between these two verbs.

(14) 我 得不到 奖学金
I get-not-arrive scholarship
‘I can**not** get a scholarship’

In Chinese, we defined 10 *core* negation markers, these being: 不, 没, 没有, 未, 别, 莫, 无, 勿, 非and 否. These core markers can be affixed to other words to create additional markers; beside adjectives, as shown in (13), it is common for markers to join adverbs, as with ‘从未’ , equivalent to the English ‘never’.

Compared to English, Chinese displays a wider inventory of sentential negation markers and the meaning of these may also differ in their distributional and aspectual constraints.

For instance, ‘没’ and its allomorph ‘没有’ carry aspectual information to deny the ‘realization’ of an event. As shown in (15), a negative sentence containing this marker is contrasted with its positive counterpart containing the aspectual particle ‘了’ flagging that an action has been completed. ‘没有’ is also one of the few examples of asymmetries between positive and negative sentences in Chinese.

(15) 我 吃饭 了
I eat ASP.
‘I have eaten’
我 **没有** 吃饭
I have not eat
‘I have **not** eaten’

Unlike English, existential constructions and imperatives also require specific markers.

The marker ‘没有’ is used with the meaning of ‘there is no/not’. Unlike its aspectual homograph ‘没(有)’, the existential marker is semantically transparent with its positive counterpart being the verb ‘有’, ‘there is’. Existential negation in Chinese is also used for universal quantification, corresponding to the English ‘no’, as shown in (16):

(16) **没有** 国家 与 马耳他 接壤
no country with Malta border
‘Malta borders no country’

Unlike English, where ‘not’ is used for the indicative and the imperative mood alike, Chinese has two dedicated markers ‘别’ and ‘勿’, which directly precede the verb they negate.

Finally, despite the word order in the two languages being similar, there are a few differences as to where the negation appears with respect to the predicate. Whereas sentential negation in English requires the adverb to directly precede the negated verb, in Chinese the verb and object order can be inverted (effectively creating an SOV order) and the latter placed in between the negation marker and the negated verb. In Chinese, this inversion is achieved through the light verb 把 as shown in (18).

(17) 我 **没** 把 他 的 钱包 偷走
I did-not BA he DE wallet steal
‘I did not steal his wallet’

Whereas negative polarity items (NPI) are present in both English and Chinese, the position of these with respect to sentential negation is different, with Chinese placing them before the negated predicate, which is not allowed in English.

(18) 我 一 点 也 没 吃
I one bit also did-not eat
‘I did not eat anything at all(?Anything at all I did not eat)’

2.4 Decomposing negation

Let us go back to the example from Ch. 1, which is shown below for convenience.

(19) Cats do not like pizza

1. $\neg\exists x.e.y.cat(x) \wedge like(e) \wedge pizza(y) \wedge arg0(e,x) \wedge arg1(e,y)$
2. Cats do **not** like pizza

When discussing negation in the previous sections, we have often referred to the presence of a negation marker and the effect it has on the surrounding context. This is straightforward in a logical form; in (19) we can identify a negation operator and the set of predicates and variables (as well as other operators) that falls in its scope.

But how does this translate to a string? Beside the negation marker, how do we represent the other components of negation? And what are these components anyway? We try here to give a clearer definition of a few key sub-components that make up negation by looking at previous work. These are mainly three:

- **Cue**: the word (e.g. 'not'), morpheme (e.g. 'im-'patient) or multi-word unit (e.g. 'no longer') inherently expressing negation. Often there is a direct mapping between a cue word and the negation operator; however, morphological negation is often not considered as negation at level of logical form. Moreover, there are words such as 'other/another' that are not considered as cue words but can be represented as containing negation in a logical form ($:= \exists x.y.p(x) \wedge p(y) \wedge \neg eq(x,y)$).

- **Scope**: There is no definitive explanation of what negation scope is at a string-level, with Morante and Daelemans (2012) defining it as "the part of the sentence affected by a negation cue" and Blanco and Moldovan (2011) defining it as the "set of elements whose individual falsity would make the negated statement strictly true". In the corpora we will be working with, negation scope is often taken to be a negated 'meaningful syntactical unit'; perhaps for this reason, guidelines mention the 'it is not the case that...' test to identify which part of a sentence belongs to the scope of negation. This means for instance that in (19.2) 'it is not the case that cats like pizza' is the same as saying 'Cats do not like pizza' and therefore we include the entire clause in the scope; but if we add a coordinate clause to yield the sentence 'Cats do not like pizza and we all know it', 'it is not the case that cats like pizza and we all know it' does not hold the same meaning, which is why we exclude the second coordinate clause from the scope. However, a syntactic definition of scope has nothing to do with truth value checking (as defined in Blanco and Moldovan (2011)) nor it is mentioned in Morante and Daelemans (2012)'s definition.

  Moreover, considering that the creation of several different corpora tailored to different domains leads to different annotation guidelines, all claiming that what was annotated is the 'linguistic scope' of negation, there is a need to systematically understand what exactly falls inside the scope of negation and what does not and how this differs from negation represented in a logical form, both problems we look at in detail in §2.5 and Ch. 4.

- **Event**: the element in the scope the cue directly refers to (e.g. 'He is **not** *driving* a car'). Even if there might be different interpretations of what a negation event is (e.g. a hierarchy of semantic events – for instance 'He is not driving a car' involves the events of 'moving', 'driving' and 'car-driving'), the event is considered as the *minimal* lexical unit that is directly negated. The term 'minimal' means that in the case of complex verbal constructions, only the head of the predicate is considered as the event (in the example above, only 'driving' is the event, not 'is driving'). Finally, not all negation instances at the string level contain an event (e.g. interjections as in 'Do you want to buy it? **No**, I do not' or in cases of verbal ellipsis 'She swims but I do **not**').

- **Focus**: the part of the scope that is directly negated or emphasized. For instance, in the sentence 'He does not want to go to school by car', the speaker might emphasize the fact that 'He does not want to go *to school* by car' (but going somewhere else by car is fine) or that 'He does not want to go to school *by car*' (but by other means of transportation). The focus is the most difficult part to deal with because of its ambiguity, with its interpretation often requiring information beyond the sentence level.

It can be inferred from some definitions, especially in the case of scope and focus, that there is not a one-to-one correspondence between these sub-components and their logical counterpart, and the terminology might be different as well. For instance, negation focus seems to correspond roughly to narrow-scope as opposed to wide-scope; for instance, in the 'He does not go to school *by car*', where *by car* is the focus, the sentence has the wide-scope reading in (15.1) as well as a narrow-scope reading in (15.2), the latter corresponding to the focus:

(20)
1. $\neg\exists e.x.y.z.go(e) \wedge he(x) \wedge school(y) \wedge car(z) \wedge arg0(e,x) \wedge to(e,y) \wedge by(e,z)$
2. $\exists e.x.y.go(e) \wedge he(x) \wedge school(y) \wedge arg0(e,x) \wedge to(e,y) \wedge \neg\exists z.car(z) \wedge by(e,z)$

Negation scope in particular is problematic since it interacts with a variety of other semantic phenomena (e.g. modality, quantification); are these interactions captured when annotating negation scope at a string level? In the next section, we will try to answer these questions by surveying existing corpora annotated for negation scope along with their guidelines.

2.5 A survey of corpora annotated for negation scope

The following are the corpora annotated for negation scope currently available (related statistics reported in Table 2.1), along with an annotated example to illustrate the genre of each (cue in **bold** and scope underlined).

- BioScope (Vincze et al., 2008, 2011) is a collection of medical and biological texts annotated for negation, speculation and their linguistic scope. The corpus consists of three subcorpora: abstracts of medical papers(*abstracts*), full papers (*full*) and clinical reports (*clinical*).

  (21) It helps activation , **not** inhibition of ibrf1 cells .

- SFU product review corpus (Konstantinova et al., 2012) is a collection of movie, book, and consumer product reviews from the website Epinions.com annotated for negation, speculation and their scope.

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

- ConanDoyleNeg (Morante and Daelemans, 2012) consists of four stories from Conan Doyle's 'Sherlock Holmes' annotated for negation event, scope and focus.

  (23) I repeat that the lady is his wife and **not** his sister

- CNeSp (Zou et al., 2016): a collection of heterogeneous texts including scientific literature, product reviews, and financial articles, annotated for both negation and speculation and their scope.

  (24) 酒店 不能 多 给 我们 提供 一个 枕头
  hotel cannot more give we offer one pillow
  'The hotel could not offer us one more pillow'

The first line we draw in between these corpora separates the first three from ConanDoyleNeg. Although the general goal is to annotate 'linguistic scope', what is

	# sentences	# negated sentences	cue	scope	event	focus
BioScope - abstract	11871	1596(13.45%)	✓	✓		
BioScope - full	2670	339(12.70%)	✓	✓		
BioScope - clinical	6383	865(13.55%)	✓	✓		
SFU	17,263	3124(18.1%)	✓	✓		
Conan-Doyle-neg.	5520	1227(22.22%)	✓	✓	✓	✓
CNeSp - scientific	4630	611(13.2%)	✓	✓		
CNeSp - product	4998	2643(52.9%)	✓	✓		
CNesp - financial	7213	1262(17.5%)	✓	✓		

Table 2.1: Statistics of the corpora annotated for negation

really annotated in the BIOSCOPE, SFU and CNESP is *syntactic scope* with a particular emphasis on easing the extraction of negated keywords, whereas the focus in the more recent CONANDOYLENEG seems to shift toward a *semantic* notion of negation scope. However, *semantic* here means that what the scope should represent is the *argument structure around a negated event* rather than mirroring a FOL representation of scope at string level.

A common feature of syntactic scope is the requirement for the scope to be a *continuous* span of text; in cases of ellipsis, such as (25), only the cue is annotated because the material omitted already takes part in an affirmative construction and it is not repeated.

(25) This decrease was seen in patients who responded to the therapy as well as in those who did **not**

Another feature of these three corpora is the treatment of the subject. The BIOSCOPE and the SFU consider scope as the maximal syntactic unit to the right of the cue; the term *maximal* specifies that given a negated constituent, all words in its yield fall inside the scope (hence ensuring continuity). This also means that the subject is usually excluded from the scope of negation, with the only exception of passive construction where the subject is annotated, with the explanation that it is the object of its correspondent active construction; this contrast is exemplified in (26)

(26) 1. A small amount of adenopathy **cannot be** completely **excluded**

2. Once again, the Disorder module does **not** contribute positively with the prediction

The exclusion of the subject seems however to go against the very purpose of detecting scope for the purpose of information extraction, where the main goal task is to detect all those elements affected by negation, subject included. For this reason, the CNeSp follows closely the BioScope annotation guidelines with the only exception of the subject which is always annotated, because it is part of the negated meaning of the sentence.

(27) 卫浴设备 **不能** 正常工作
Shower facilities could-not often work
Shower facilities couldn't work often

On the other hand, in the annotations of ConanDoyleNeg the scope can be a *discontinuous* span of text to include more fine-grained linguistic phenomena, such as ellipsis and long-range dependencies. Related descriptions of how negation scope at a string-level should look like often mention the importance of argument structure, that is to identify all arguments around a negated statement (Blanco and Moldovan, 2011). In the example in (28), 'She swims but I do not', we would therefore detect as part of negation scope the event, 'swims' as well as the subject 'I'. To do this, a discontinuous scope helps us to annotate those unrepeated, omitted elements that appear only in the affirmative clause.

(28) She swims but I do **not**

ConanDoyleNeg also extends the inventory of negation cues considered by including affixal negation. As shown in 29, this introduces cases of NP negation in the annotations.

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

Once again, we underline that ConanDoyleNeg introduces a more semantic notion of negation scope as argument structure to improve on a purely a syntactic (and sometimes questionable) annotation of negation scope. There are however differences between this definition and logic negation understood in terms of truth values, despite the lines being sometimes blurred. Specifically, string-level negation presents the following limitations:

The status of quantifiers. The first of these concerns the status of quantifiers. The set notion behind existential quantifiers like *some* and *a few* as well as universal quantification triggered by words such as *every*, *each* or *all* is inevitably lost when projecting to strings. If that is less of a problem for existential quantifiers, it is a limitation for universal quantifiers. This is exemplified in (30) where we show the ConanDoyleNeg annotations against the semantic interpretation.

(30) i Some students have **not** done the homework.
$\neg\forall x.student(x) \rightarrow \exists e.y.do(e) \wedge homework(y) \wedge arg0(e,x) \wedge arg1(e,y)$
$=$ It is not the case that all students have done the homework (some have, some have not).

ii All students have **not** done any homework
$\forall x.student(x) \rightarrow \neg\exists e.y.do(e) \wedge homework(y) \wedge arg0(e,x) \wedge arg1(e,y)$
$\neq$ It is not the case that all students have done the homework (some might have)

iii **Not** all students have done the homework
$\neg\forall x.student(x) \rightarrow \exists .e.y.do(e) \wedge homework(y) \wedge arg0(e,x) \wedge arg1(e,y)$
$=$ It is not the case that all students have done the homeworks (some might have)

Whereas in (30.i), the string-level annotations mirror the logical form, the same doesn't hold for (30.ii) where the annotations at a string level cannot distinguish from the different interactions between the scope of the universal quantifier and the scope of negation.

Ellipsis. Despite the ConanDoyleNeg corpus annotating ellipsis, there is no way to annotate an omitted event as part of negation scope and outside of it at the same time. This is the case of ex. 28 where 'swims' is marked in the scope of negation despite not being such in the first coordinated clause.

Implicit negation. Given that the presence of the cue defines the presence of negation, when this is not explicit at a string level, its scope cannot be captured. As shown above, this includes words such as 'another' or 'other' as well as to expressions denoting set membership. For instance in the following sentence,

(31) Only the people in class A went to party

'only' acts a restrictor on the set of people going to the party, where there is no one from other classes attending the event, and therefore having an implicit negative meaning.

2.5.1 What is considered as negation

In order to clarify what is to be considered negation in this thesis, and what is not, the following section will summarize the various criteria that we have used to include and exclude specific kinds of negation in line with the investigative aims of this thesis. Note that we base most of the criteria on the available annotations in ConanDoyleNeg. We will return to these criteria in Ch. 4 when describing the annotations in English and Chinese in greater detail.

In general we consider as negation instances where an explicit cue is present and the cue carries a clear negative meaning. This includes:

Sentential negation introduced by lexical cues. In English this is generally introduced by 'not' or 'never'. This also includes discontinuous negation (e.g. 'neither...nor').

Non-sentential negation introduced by lexical cues, where the scope spans a secondary constituent only. In English this includes, amongst others, the cues 'without' and 'except'.

Negated quantifiers introduced by lexical cues (such as 'not many' or 'no') as well as by pronoun-cues such as 'nobody', 'nothing' and 'nowhere' in any syntactic environment.

Affixal negation, introduced by morphological cues such as 'im-' (e.g. 'impatient') or 'un-' ('uncooked'). However, the status of affixal negation as carrying contradictory meaning is disputable: these constructions are in fact 'weak contraries', where the 'Law of Excluded Middle' (where either a statement or its negation has to be true) does not apply. For instance, whereas the statement "Sam eats bread or it is not the case that Sam eats bread" is true given that there is not third option in which Sam either eats bread and he also does not, "Sam is happy or it is the case that he is unhappy", does allow for non-excluded middle where Sam can be somewhat happy. [1]

What is **not** considered as negation:

Implicit negation. In general we exclude all instances where there is not an explicit

[1] The ConanDoyleNeg annotations are not consistent in annotating affixal negation on adverbs. Despite the guidelines stating that when an adverb is negated it takes wide-scope, where the entire clause should be place under the scope of negation (as opposed to the *focus* where the scope should be narrow, i.e. on the adverb alone), annotations are not consistent, as shown below.

(32)
1. Far away on the path, we saw Sir Henry looking back, [...] glaring hope**lessly** at the frightful thing which was hunting him down .
2. I found myself [...] tossing rest**lessly** from side to side.

We will return to this point in Ch. 4.

cue marking negation. Some of these instances we have discussed in the previous section; for instance we do not consider ‘another’ or ‘other’ as carrying the meaning of ‘an entity which is not the same as some other’. In the same way, we do not consider ‘inferred negation’, where a negative meaning can be extrapolated from a positive statement (e.g. ‘The bag is too heavy too lift’ implies that ‘The bag cannot be lifted’).

Non-functional negation. We do not consider as part of our work cues that carry a positive meaning. This includes, amongst others:

- Question tags: we consider question tags (e.g. ‘It is interesting, *isn’t it*?’) as not conveying a real negative meaning since their purpose is mostly to seek confirmation from the addressee.

- Fixed expressions: expressions like ‘cannot help but’ and ‘none the less’ despite containing a negation cue are not considered as carrying negative meaning and therefore are not considered as negation. Amongst one of the most recurrent fixed expressions, we also exclude ‘not only’ since it carries a positive meaning (e.g. ‘Not only James, but also Sarah went to the party’ = ‘Both Sarah and James went to the party’).

- Modality cues: we exclude expressions like ‘no doubt’ or ‘without a doubt’ which indicates a strong degree of certainty that an event has indeed happened.

2.6 Detecting negation scope: a survey of previous work

Given the availability of data annotated for negation, it is natural that previous work has cast the problem of negation scope detection as a supervised machine-learning task.

Previous work can be divided into three main categories according to the type of system proposed: *rule-based*, *machine-learning* and *hybrid systems*.

Rule-based systems prove to be a valid choice when either the domain is narrow and well-defined or when one wants to capture the syntactic regularities of how negation is annotated in some of the corpora. Early negation scope detection systems, such as NegEx (Chapman et al., 2001) exploit a database of relevant in-domain terminology (in this case, medical) along with a hand-crafted list of regular expressions to capture the scope of negation from clinical records.

People have also exploited the underlying syntax of negation scope and have created rules to detect the scope of negation in a constituency tree. Jia et al. (2009) have shown that detecting scope via constituency tree-traversal heuristics helps in detecting the sentiment in documents from the microblog domain; Ballesteros et al. (2012) have also proposed a similar approach but evaluated intrinsically on the CONAN-DOYLE-NEG corpus, while de Albornoz et al. (2012) use a dependency rather than a constituency tree.

There exists also approaches that tried to detect negation scope at a string level by leveraging its logical form. Among these, Basile et al. (2012) investigated whether one can map negation scope in the form of Discourse Representation Graph (DRG) to the actual tokens in the sentence.

The vast majority of previous work however, takes advantage of the fact that annotations are available to train a supervised machine-learning systems. Models for automatic negation scope detection treat the task as a tagging problem; what is tagged differs according to the approach chosen. Sequence classifiers tags each tokens in a sentence as being either inside or outside negation scope; structured classifiers classify whether a syntactic constituent is in the scope of negation or not.

Morante et al. (2008) and Morante and Daelemans (2009) propose a series a sequence classifiers that classify the scope in two steps: first, they identify the negation cue and based on this, they then identify the elements in the scope in IOB (Inside Outside Begin) fashion. Beside experimenting with Support Vector Machines(SVM) and Conditional Random Fields(CRF) separately, they also show that by using another classifier as a meta-learner that combines the output of these two models, they achieve the best performance on the BIOSCOPE corpus. Councill et al. (2010) has also proposed using CRF to detect negation scope; unlike Morante et al. (2008) however, cues are detected from a hand-built list of 35 elements and scope resolution leverages features extracted from a parse tree. The importance of dependency-based features in sequence classifiers has also been recognized by Lapponi et al. (2012) which yield the best performance amongst the purely machine-learning based systems submitted for the *SEM2012 shared task on automatically detecting negation scope.

There are then several hybrid models that combine the use of heuristics with machine learning models. Li et al. (2010) propose a multi-pass machine-learning model to classify constituents in the tree as being part of negation scope. In a first pass, candidates that could potentially be in the scope of negation are proposed via a set of simple heuristics, which are subsequently filtered in a second pass using an SVM classifier.

Read et al. (2012) also propose a similar approach where a set of tree traversal rules are first manually built and then reranked by an SVM classifier.

Similarly to Basile et al. (2012), Packard et al. (2014) have investigated if starting from an formal meaning representation is beneficial for the task of negation scope detection. They do so by leveraging Minimal Recursion Semantics (Copestake et al., 2005); semantic representations are first extracted using the Lingo English Resource Grammar (Flickinger, 1999) and in a second step mapped to the sentence through hand-crafted heuristics. If for a given sentence a parse is not found, the systems backs-off to the syntactic-based ranker of Read et al. (2012).

Work in detecting negation in Chinese doesn't differ from the aforementioned methods. Zou et al. (2016) uses a syntactic method very similar to Li et al. (2010), except that dependency parse are used in place of constituency trees. Again, ancestors of the cue in a constituent tree are considered as potential roots for negation scopes and reranked according to a feature classifier.

What works and what doesn't? In general, rule-based methods lack flexibility and unsurprisingly, they are often outperformed by machine-learning classifiers. Rules usually define scope as the sub-tree dominated by the parent of the negation cue; if this correctly identifies negation on the main predicate (assuming gold trees), it needs an additional set of rules to handle different syntactic environments, which are often incorrectly detected.

Among machine-learning methods, sequence classifiers in English show a better performance than classifiers purely based on syntactic information. However, adding syntactic features in the form of dependency or constituency paths from the cue to a given token seem to help, as well as adding post-processing heuristics to correct the classifier output.

Methods that map negation from a logical form to the string-level to detect its scope do not perform as well as other machine-learning methods, unless it is combined with other systems (as in the case of Packard et al. (2014)). This is due to two main reasons: first, the logical form abstracts from certain elements that are instead part of the scope at a string level, such as determiners and other function words; these need to be recovered separately. Second, the semantic parser producing the logical form might fail to output a parse at all.

Chapter 3

Recurrent neural networks for negation scope detection

The content of this chapter is based on two published, peer-reviewed papers:

- Fancellu, F., Lopez, A. & Webber, B., *Neural Networks For Negation Scope Detection.* in The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Association for Computational Linguistics, pp. 495-504

- Fancellu, F.; Lopez, A,; Webber, B.; He, H., *Detecting negation scope is easy, except when it isn't.* Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics (ACL), 2017. p. 58-63.

3.1 Introduction

In Chapter 2, we have shown that previous work have tackled the problem of detecting negation scope in Chinese by using similar techniques for English. The fact that similar techniques are used is not surprising given that the formulation of the task is the same in both languages: the definition of negation scope in the only corpus available in Chinese, the CNESP, is in fact heavily based on the one formulated for the BIOSCOPE corpus for English.

However, despite this similarity, none of the previous work has investigated whether or not there exists a model that can be generalized to both languages. We then ask the following question, *can we then build a classifier that exploits the same set of* *features regardless of the language, while performing as good as or better than previous approaches?*

To answer this question we investigate whether neural network based classifiers are a valid alternative for the task. The first advantage of neural networks-based methods is that we can perform classification by means of (a few) unsupervised embedding features only, easy to obtain for languages other than English with the proviso that a large corpus of text is available. A second advantage is in the ability of some of these models to store memory of previous time steps; this can be advantageous in cases of discontinuous scope.

We start by testing our approach on the CONANDOYLENEG corpus, where negation scope detection is more challenging due to the more fine-grained annotation guidelines. We only consider two features to detect negation scope, word and PoS embeddings; while words encode token-specific information, we hypothesize PoS information to encode shallow syntactic information, while being easy to extract for a large number of languages.

Results show that bi-directional Long-Short Term Memory Networks (BiLSTM below) outperform previously developed classifiers in presence only of two embedding features, word and universal PoS tags.

Given these positive results, we then tested the robustness by answering two different questions:

1. is our model still state-of-the-art when tested on corpora of a different genre?
2. does the same approach work equally well across all available corpora annotated for negation scope?

Considering 1., we experimented with training the model on CONANDOYLENEG and testing on annotated sentences from Simple Wikipedia. We found that our model sometimes fails to outperform previously developed non-neural classifiers that also exploit syntactic features. We also found that performance is lower when recognizing scope around morphological negation.

As for 2., we extend our method by training and testing on all other available corpora annotated for negation scope in English and Chinese. Experiments on the Chinese corpus are of particular importance to assess whether our method is generalizable across languages.

Although confirming that our model is state-of-the-art, we show that in most corpora, negation scope is often delimited by punctuation or sentence boundaries. That is, in these corpora, examples like (33) outnumber those like (34).

(33) It helps activation , **not** inhibition of ibrf1 cells .

(34) She swims but I do **not**

Our experiments demonstrate that negation scope detection is very accurate for sentences like (33) (which we call the ‘easy’ case here for convenience) and poor for others (the ‘hard’ cases), suggesting that most classifiers simply overfit to this feature of the data. When we attempt to mitigate this effect by under-sampling the ‘easy’ cases during training, our system does not improve on the ‘hard’ ones, suggesting that more training data is required to make progress on the phenomena they represent.

3.2 The task

We begin formalizing the task of detecting negation scope by giving some definitions. A *negative sentence n* is defined as a vector of words $\langle w_1, w_2 ... w_n \rangle$ containing one or more negation cues, which we have already defined to be a word (e.g. *not*), a morpheme (e.g. *im*-patient) or a multi-word expression (e.g. *by no means, no longer*) inherently expressing negation.

A word is a scope token if included in the scope of a negation cue. Each cue defines its own *negation instance*, here defined as a tuple $I(n,\mathrm{c})$ where $c \in \{1,0\}^{|n|}$ is a vector of length n s.t. $c_i = 1$ if w_i is part of the cue and 0 otherwise. Given I the goal of automatic scope detection is to predict a vector $s \in \{0,1\}^{|n|}$ s.t. $s_i = 1$ (inside of the scope) if w_i is in the scope of the cue or 0 (outside) otherwise.

In (35) for instance, there are two cues, *not* and *no longer*, each one defining a separate negation instance, *I1(n,c1)* and *I2(n,c2)*, and each with its own scope, *s1* and *s2*. In both (35a) and (35b), n = [I, do, not, love, you, and, you, are, no, longer, invited]; in (35a), the vector *c1* is 1 only at index 3 (w_3=‘not’), while in (35b) *c2* is 1 at position 9, 10 (where w_9 w_{10} = ‘no longer’); finally the vectors *s1* and *s2* are 1 only at the indices of the words underlined and 0 anywhere else.

(35) a. I do **not** love you and you are no longer invited.

b. I do not love you and you are **no longer** invited

There are the two main challenges involved in detecting the scope of negation in ConanDoyleNeg: 1) a sentence can contain multiple instances of negation, sometimes nested and 2) scope can be discontinuous. As for 1), the classifier must correctly classify each word as being inside or outside the scope and assign each word to the correct scope; in (36) for instance, there are two negation cues and therefore two scopes, one spanning the entire sentence (3a.) and the other the subordinate only (3b.), with the latter being nested in the former (given that, according to the ConanDoyleNeg guidelines, if we negate the event in the main, we also negate the whole main clause).

(36) a. I did **not** drive to school because my wife was not feeling well .[1]

b. I did not drive to school because my wife was **not** feeling well .

In (37), the classifier should instead be able to capture the long range dependency between the subject and its negated predicate, while excluding the positive VP in the middle.

(37) Naomi went to visit her parents to give them a special gift for their anniversary but **never** came back .

Finally, negation scope can sometimes be empty as in the case of interjection 'No'.

(38) **No**, sir; it is gone forever.

3.3 The model

Long-Short Term Memory Networks (LSTM) are recurrent neural models where each cell regulates the flow of information coming from the input and from previous time-steps through different gates, which, unlike RNNs, can better retain information over long-sequences. Bi-directional LSTM (BiLSTM) are an enhancement of the basic LSTM architecture where the input is recursed in two directions, forward, from left to right, and backwards, in the opposite direction, in order to fully model the dependencies between the different parts of the input. A Bi-directional LSTM architecture is a perfect fit for the task of negation scope detection for two main reasons: *bi-directionality* is essential given that a scope token can appear in a string both before and after the cue

[1] One might object that the scope only spans over the subordinate given that it is the part of the scope most likely to be interpreted as false (*It is not the case that I drove to school because my wife was not at home, but for other reasons*). However, in ConanDoyleNeg, this is defined as the 'focus' of negation and considered as part of a wide scope. We will come back to this and related issues in the annotations in Ch 4

and the model needs to be aware that the latter is present before starting to detect scope; *long-short term memory* is important given that scope might span a long sequence and be discontinuous.

In our model, the BiLSTM takes as input a single negative instance $I(n,c)$. We represent each word $w_i \in n$ as a d-dimensional word-embedding vector $\mathbf{w} \in \mathbb{R}^d$. In order to encode information about the cue, each word is also represented by a *cue*-embedding vector $\mathbf{c}$, which is a binary flag in that it can only take two representations, 'cue', if c_i=1, or 'notcue' otherwise. We also define $\mathbf{E}_w^{vxd}$ as the word-embedding matrix, where v is the vocabulary size, and $\mathbf{E}_c^{2xd}$ as the cue-embedding matrix. Additionally we experiment with a third embedding vector $\mathbf{p}_i$, encoding information about the PoS (or universal PoS tag) of word w_i, with associated PoS-embedding matrix $\mathbf{E}_p^{vxd}$.

The input to the network for each word w_i are the word-embedding vector $\mathbf{w}_i$, the cue-embedding vector $\mathbf{c}_i$ and optionally a PoS-embedding vector $\mathbf{p}_i$, which we concatenate together to form the input vector $\mathbf{x}_i$. The computation of the hidden layer for time step t and the output can be represented as follows:

$$\begin{aligned}
\mathbf{x}_t &= [\mathbf{w};\mathbf{c}(;\mathbf{p})] \\
\mathbf{i}_t &= \sigma(\mathbf{W}_x^{(i)}\mathbf{x} + \mathbf{W}_h^{(i)}\mathbf{h}_{t-1} + \mathbf{b}^{(i)}) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_x^{(f)}\mathbf{x} + \mathbf{W}_h^{(f)}\mathbf{h}_{t-1} + \mathbf{b}^{(f)}) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_x^{(o)}\mathbf{x} + \mathbf{W}_h^{(o)}\mathbf{h}_{t-1} + \mathbf{b}^{(o)}) \\
\tilde{\mathbf{c}}_t &= tanh(\mathbf{W}_x^{(c)}\mathbf{x} + \mathbf{W}_h^{(c)}\mathbf{h}_{t-1} + \mathbf{b}^{(c)}) \\
\mathbf{c}_t &= \mathbf{f}_t \cdot \tilde{\mathbf{c}}_{t-1} + \mathbf{i}_t \cdot \tilde{\mathbf{c}}_t \\
\mathbf{h}_{back/forw} &= \mathbf{o}_t \cdot tanh(\mathbf{c}_t) \\
y_t &= g(\mathbf{W}_y([\mathbf{h}_{back};\mathbf{h}_{forw}]) + \mathbf{b}_y)
\end{aligned} \tag{3.1}$$

where $\mathbf{W}_x$ and $\mathbf{W}_h$ are matrices that weigh the input and the hidden state of the previous time step respectively, $\mathbf{h}_{t-1}$ the hidden layer state a time t-1, $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ the input, forget and the output gate at the time t and [$\mathbf{h}_{back}$; $\mathbf{h}_{forw}$] the concatenation of the backward and forward hidden layers.

Finally, the training objective is to maximize, for each negative instance, the negative log likelihood *J(W,b)* of the correct predictions over gold standard labels:

$$\begin{aligned}
J(W,b) = -\frac{1}{l}\sum_{i=1}^{l} y^{(w_i)} \log h_\theta(x^{(w_i)}) \\
+ (1-y^{(w_i)})\log(1-h_\theta(x^{(w_i)}))
\end{aligned} \tag{3.2}$$

where l is the length of the sentence $n \in I$, $\mathrm{x}^{(w_i)}$ the probability for the word w_i to belong to either the I or O class and $y^{(w_i)}$ its gold label.

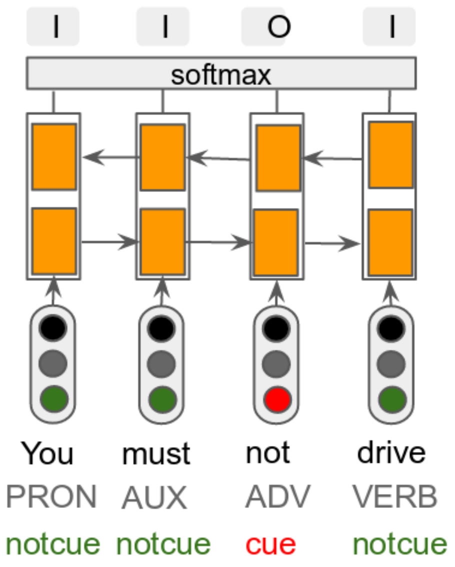


Figure 3.1: BiLSTM architecture for negation scope detection. Each token in a negated sentence is represented by the concatenation of word, PoS and cue embedding features. The cue embedding feature is a binary flag, here represented in red if the word is a cue, green otherwise.

The BiLSTM architecture is shown in Figure 3.1.

3.4 Experiments

We start by training and testing on the CONANDOYLENEG dataset. The training, dev and test sets are of 984, 173, 264 sentences respectively. These are all negative sentences only, i.e. those sentences with at least one cue annotated. If a sentence contains multiple negation instances, we create as many copies as the number of instances.

Our *vanilla model*(BiLSTM) uses cue and word embeddings as features, where the embeddings are randomly initialized.

Additionally, we experimented with the following settings:

1. *Token-distance baseline*: In order to understand how difficult the task of negation scope detection is, we created a simple baseline by tagging as part of the scope all the tokens 3 words to the left and 6 to the right of the cue; these values were found to be the average span of the scope in either direction in the development data.

2. *Sentence baseline*: As an additional baseline, we consider all the words in the sentence to be in the scope of negation.

3. *Pre-trained embeddings (+E)*: instead of randomly initialized embeddings, we use pre-trained word embeddings. We hypothesize that given the small size of the training data using pre-trained word embeddings might boost the performance of the classifier. We experimented both with keeping the word-embedding matrix fixed and with updating it during training but we found no difference between the two settings; again, this might be due to the size of the training data, too small to properly tune the pre-trained embeddings. We train a word-embedding matrix using Word2Vec (Mikolov et al., 2013b) on 770 million tokens (for a total of 30 million sentences and 791028 types) from the 'One Billion Words Language Modelling' dataset [2] and the Sherlock Holmes data set combined.

4. *Adding PoS / Universal PoS information (+PoS/+uni PoS)*: This was mainly to assess whether or not we could get further improvement by adding shallow syntactic information. In all the settings above, we also add an extra embedding input vector for the PoS or Universal PoS of each word w_i. In order to maintain consistency with the original data, we perform PoS tagging using the GENIA tagger (Tsuruoka and Tsujii, 2005)[3] and then map the resulting tags to universal POS tags.[4]

5. *Segmenting morphological negation* : If the sentence contains a morphological cue (e.g. **im-**patient) we also experimented with splitting it into affix (**im-**) and root ('patient'), and consider the former as cue and the latter as part of the scope. When using pre-trained word-embeddings, we also tokenise the corpus by splitting a word containing morphological negation into negation affix and root (e.g. 'impatient' → 'im-' and 'patient') to match the CONANDOYLENEG corpus. In order to perform this split, we matched each word against a hand-crafted list of words containing affixal negation[5]; this method has an accuracy of 93% on the Conan Doyle test data.

   We compare segmenting morphological negation vs. not segmenting it to see whether it leads to any difference in the performance of the classifier.

[2] Available at `https://code.google.com/archive/p/word2vec/`

[3] `https://github.com/saffsd/geniatagger`

[4] Mapping available at `https://github.com/slavpetrov/universal-pos-tags`

[5] The list was courtesy of Ulf Hermjakob and Nathan Schneider.

Both neural network architectures are implemented using DyNet (Neubig et al., 2017) and the Adam optimizer (Kingma and Ba, 2014) with a starting learning rate of 0.001 and a dropout rate of 0.2 on the output layer. We tune the hyperparameters on our basic setting (i.e. using cue and word embeddings features only) using randomized grid-search (d_h=200, d_w=100, d_p=100,d_c=20). We use these for all our models and baselines. We report results as an average on 5 runs. Statistical significance across settings is calculated using t-test.

3.5 Evaluation

We evaluate our classifier in two different ways. First, we measure token-level precision, recall and F_1 over tokens identified as within scope. Second, we compute the *percentage of correct scope* (PCS), as precision, recall and F_1 measure over scopes that we fully and exactly match in the test corpus. In order to compare the performance of our model to previous work, when computing the PCS, we follow the definition given by Morante and Blanco (2012): false positive includes those scopes that we predict but are absent in the gold annotation (e.g. (38)) while false negatives are those scopes that we fail to predict but are instead present in the gold standard. Those scopes that we only partially predict are included in the false negative counts only (hence, impacting recall).

Results on the held out set are compared to two different systems: the best system from the *SEM2012 shared task, UiO1 (Read et al., 2012, for details we refer the reader to section §2.6), and Packard et al. (2014), an hybrid rule-based/ machine-learning system where negation is detected using a formal semantic representation, MRS, along with a re-ranker.

3.6 Results

The results of the scope detection task are shown in Tables 3.1 and 3.2. We found the addition of shallow syntactic information in the form of PoS to help the performance of our system; this improvement over a vanilla system relying on cue information only is statistically significant at p<0.001. Results also show that there is no difference in using PoS and universal PoS tags; the difference between the two settings is not statistically significant (p=0.36). This is an important result since **utilizing universal PoS allows for a truly generalizable system with features which are the same irrespective of the language**.

	Scope tokens			Exact scope match		
System	*Prec.*	*Rec.*	F_1	*Prec.*	*Rec.*	F_1
TD-Baseline	66.23	22.37	33.44	20	0.006	0.01
S-Baseline	38.08	100	55.16	0.0	0.0	0.0
BiLSTM	92.71	81.87	86.93(± 0.37)	100	64.16	78.17(±0.84)
BiLSTM (+ PoS)	93.24	**84.66**	**88.70**(± 0.45)	100	63.58	77.74(± 0.68)
BiLSTM (+ UniPos)	**93.65**	83.64	88.63(± 0.4)	100	**66.47**	**79.86**(± 0.97)
BiLSTM (+ E)	93.07	80.82	86.51(± 0.39)	100	61.85	76.43(± 0.83)
BiLSTM (+ PoS + E)	92.62	82.21	87.09(± 0.52)	100	57.80	73.26(± 1.27)
BiLSTM (+ UniPoS + E)	92.49	82.63	87.28(± 0.3)	100	64.16	78.17(± 0.73)

Table 3.1: Results for the scope detection task on the *dev* set. Results are compared against the token-distance baseline (TD-Baseline) and the sentence baseline (S-baseline). Standard deviation are reported for the F_1 measure.

	Scope tokens			Exact scope match		
System	*Prec.*	*Rec.*	F_1	*Prec.*	*Rec.*	F_1
TD-Baseline	58.60	22.91	32.94	18.75	0	0.02
S-Baseline	31.99	100	48.48	0	0	0
Best closed track: UiO1	81.99	88.81	85.26	87.43	61.45	72.17
Packard et al. (2014)	86.1	**90.4**	88.2	98.8	65.5	78.7
BiLSTM (+ Uni Pos)	93.15	**86.72**	**89.82**	**98.91**	**69.47**	**81.61**

Table 3.2: Results for the scope detection task on the *test* set.

On the other hand pre-trained word embeddings do not lead to an improvement when compared to randomly initialized word-embeddings; the difference between the two settings are significant at $p<0.01$.

Finally, we didn't find any difference between segmenting words containing morphological negation into affixal cue and root and keeping the word as is; the difference in performance on the dev set is in fact not statistically significant(p=0.36).

Table 3.2 also shows that our system is able to outperform previous work in all cases except for token-based recall, even in absence of any syntactic or hand-crafted heuristics (as in Packard et al. (2014)). We also observe a better performance on the test set when compared to the development set.

3.7 Error analysis

In order to understand the kind of errors our best classifier makes, we performed an error analysis on the development set, using the BiLSTM-C+uniPoS as classifier. By analyzing the 44 scopes that the system does not predict fully, we found the following error patterns (system predictions are marked in curly brackets {}):

- in 6 cases, the system could not predict spans inside a discontinuous scope, which usually reflect long-range syntactic dependencies. As shown in (39), the system fails to retrieve material inside the adjectival phrase which is consider inside the scope of NP negation

  (39) [...] who unite in their fear and {**dis**like of their master}

  That doesn't mean that our system always fails at capturing discontinuous scope; as shown below our system is sometimes able to retrieve long-range dependencies associated with the scope of negation, as in the case of elided subject in coordinated constructions

  (40) [...] {he would} lie low and {make} **no** {move so long as he thought he was in any danger}

- in 5 cases, the system makes an incorrect prediction in presence of coordination structures; we hypothesize this is due to the fact that the system cannot distinguish between VP and NP conjunction. As shows in (41), the system fails to include part of the subject or does not include part of the object

  (41) 1. I think, Watson, a brandy and {soda would do him} **no** {harm}
  2. {It was} suggested, but **never** {proved, that the deceased gentleman may have had valuable in the house}, and that their abstraction was the motive of the crime

- we found 2 cases, where the system failed to correctly detect scope in inverted construction where the object is put in focus position, as shown in (42).

  (42) Private detectives are a class with whom {I have absolutely} **no** {sympathy}

Finally as in Packard et al. (2014), we also noticed 5 cases where the gold annotations do not follow the guidelines. Two of these concern the annotation of the scope around morphological negation on an adjective; whereas the guidelines state that the scope should span the noun phrase the adjective appears in, in (43) it is instead the entire clause to be negated. We also observed cases where a subordinate clause depending from a negated main clause is excluded from the scope whereas the guidelines specifies it should always be included; this is exemplified in (44).

(43) I have had the most singular and **un**pleasant experience [...]

(44) If it were the devil himself a constable should **never** thank God [...]

3.8 Evaluating across-genre

3.8.1 Methodology

By training and testing on the same text, we have shown that bi-directional LSTM are able to outperform previous classifiers with only two embedding features.

One question left unanswered by this as well as by previous work is whether the performance of scope detection classifiers is robust against data of a different genre and whether different types of negation lead to difference in performance. To answer this, we compare our best model with the only previously developed model we found available (White, 2012).[6]

We use here a set of sentences extracted from Simple Wikipedia manually annotated for cue and scope. The annotations were carried out by the author of this thesis according to the annotation guidelines released in concomitance with the *SEM2012 shared task (Morante et al., 2011) on automatic negation detection. We created 7 different subsets to test different types of negative sentences:

Simple: we randomly picked 50 positive sentences, containing only one predicate, no dates and no named entities, and we made them negative by adding a negation cue (*do* support or minor morphological changes were added when required). If more than a lexical negation cue fit in the context, we used them all by creating more than one negative counterpart, as shown in (45). The sentences were picked to contain different kind of predicates (verbal, existential, nominal, adjectival).

(45)
1. People talk about this topic
2. People do **not** talk about this topic

[6] Originally, White's system used automatically detected cues. In order for the results to be comparable, we feed it gold-standard cue.

3. People **never** talk about this topic

		Scope tokens			Exact scope match		
neg. type	*systems*	P	R	F_1	P	R	F_1
simple	White(2012)	100	97.65	**98.81**	100	93.98	**96.90**
	BiLSTM (+ UniPoS)	100	96.34	98.13	100	88.64	93.98
lexical	White(2012)	86.59	80.10	83.22	100	58.41	**73.75**
	BiLSTM (+ UniPoS)	87.11	83.52	**85.28**	100	48.31	65.15
prefixal	White(2012)	68.98	73.40	71.12	100	32.76	**49.35**
	BiSLTM (+ UniPoS)	69.82	72.50	**71.13**	100	22.22	36.36
suffixal	White(2012)	91.76	78	**84.32**	100	69.23	**81.82**
	BiLSTM (+ UniPoS)	95.83	54.13	69.17	100	40	57.14
multi-word	White(2012)	89.62	70.63	79.00	100	9	16.67
	BiLSTM (+ UniPoS)	92.82	71.83	**80.98**	100	40	**57.14**
unseen	White(2012)	77.53	62.73	69.35	100	38.89	**56**
	BiLSTM (+ UniPoS)	64.79	80	**71.60**	100	31.03	47.37
avg.	White(2012)	85.74	77.08	80.97	100	50.37	62.41
	BiLSTM (+ UniPoS)	85.06	76.38	79.38	100	45	59.52

Table 3.3: Performance of our best model (a bi-directional LSTM using universal PoS features) compared against White (2012)'s CRF classifier across-genre. For clarity, we boldface the F_1 measure of the best system for each subset only.

Lexical: we randomly picked 10 sentences[7] for each **lexical** (i.e. one-word) cue in training data (these are *not*, *no*, *none*, *nobody*, *never*, *without*)

Prefixal: we randomly picked 10 sentences for each prefixal cue in the training data (*un-*, *im-*, *in-*, *dis-*, *ir-*)

Suffixal: we randomly picked 10 sentences for the suffixal cue *-less*.

Multi-word: we randomly picked 10 sentences for each multi-word cue (*neither...nor*,*no longer*,*by no means*).

Unseen: we include 10 sentences for each of the negative prefixes *a-* (e.g. *a*-cyclic), *ab-* (e.g. *ab*-normal) *non-* (e.g. *non*-Communist) that are not annotated as cue in the Conan Doyle corpus, to test whether the system can generalise the classification to unseen cues.

[7] In some cases, we ended up with more than 10 examples for some cues given that some of the sentences we picked contained more than a negation instance.

3.8.2 Results

Table 3.3 shows the results for the comparison on the synthetic test set. The first thing worth noting is that by using word and universal PoS features only it is possible to reach comparable performance with a classifier using syntactic features; this is particularly evident in the multi-word and lexical sub-sets. In general, genre effects hinder our system.

Performance gets worse when dealing with morphological cues and in particular in the case of our classifier, with suffixal cues; on a closer inspection however, the cause of such poor performance is attributable to a discrepancy between the annotation guidelines, which we follow to annotate the new test data, and the annotations in the training data. Whereas the guidelines state in fact that "If the negated affix is attached to an adverb that is a complement of a verb, the negation scopes over the entire clause"(Morante et al., 2011, p. 21), 3 out of 4 examples of suffixal negation in adverbs in the training data mark the scope on the adverbial root only and that's what our classifiers learn to do.

Finally, it can be noticed that our system does worse at exact scope matching than the CRF classifier. This is because White (2012)'s CRF model is built on constituency-based features that will predict scope tokens based on constituent boundaries (which, as we said, are good indicator of scope boundaries), while neural networks, basing the prediction only on word-embedding information, might extend the prediction over these boundaries or leave 'gaps' within.

3.9 Evaluating across-corpora and languages

3.9.1 Methodology

The definition of the task is the same as that presented in §3.2. The model trained on CONANDOYLENEG makes $|w|$ independent predictions on whether word w_i is inside the scope of negation or not as determined by probability $p(s_i|w,c)$, where the dependence on w and c is modeled by encoding them using a bidirectional LSTM.

Although we show this model to be already state-of-the-art, we thought it could be further improved by modeling a dependence between the predictions of adjacent tokens. This decision was motivated by the fact that unlike CONANDOYLENEG, where the annotations allows the scope to be *discontinuous*, all other corpora defines negation scope as a *continuous* span of text. Explicitly ensuring that this span does not contain gaps seems to be necessary as shown by the prediction of the CRF classifier of Morante et al. (2008), where post-processing heuristics were applied to close these gaps.

Unlike previous work we model dependencies between predictions automatically, introducing a new joint model $p(s|w,c)$, defined as:

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

The only functional change to our previous model is the addition of a CRF to create the dependence on s_{i-1}. We use here the BiLSTM+CRF architecture of Ma and Hovy (2016), where in addition we define a class of conditional probability over all possible sequences of two consecutive labels (given that we work with 2 labels, this results in 4 possible sequences), with related weight and bias **W** adn **b**. During training, for each training instance pair i we use maximum likelihood estimation as following:

$$L(\mathbf{W},\mathbf{b}) = \sum_{i} logp(y \mid z;\mathbf{W},\mathbf{b})$$

where the goal is choose the parameters that maximize the log-likelihood of L(**W**,**b**) and where y is the sequence of output labels and z the input vectors. During both training and testing, we use the Viterbi algorithm for efficient decoding.

3.9.2 Data and Experimental Parameters

We experiment with two English corpora, the SFU product review corpus (Konstantinova et al., 2012) and the BioScope corpus (Vincze et al., 2008), and a Chinese one, the Chinese Negation and Speculation (CNeSp) corpus (Zou et al., 2016).

Statistics on the number of negation instances extracted are reported in Table 3.4. For a discussion on the annotation styles of each of these corpora, we refer the reader back to § 2.5.

Again, since we focus on scope detection, we use gold cues as input. We train and test on each corpus separately. We first extract only those sentences containing at least one negation cue and create a 70%/15%/15% split of these for training, development and test respectively. We use a fixed split in order to define a fixed development set for error analysis, but this setup precludes direct comparison to most prior work has used 10-fold cross-validation. Nevertheless, we felt a data analysis was crucial to understanding these systems, and we wanted a clear distinction between test (for reporting results) and development (for analysis).

Data		train	dev	test
SFU		2450	525	525
BioScope	Abstract	1190	275	275
	Full	210	45	45
	Clinical	560	120	120
CNeSp	Product	2744	588	588
	Financial	1053	241	241
	Scientific	109	22	22

Table 3.4: Number of the training, dev. and test instances for the corpora used in the present work.

Model parameters and initialization are the same as the ones described in § 3.4. We pretrain our Chinese word embeddings on Wikipedia data[8] and segment Chinese words using the NLPIR toolkit[9] . For Chinese, we experimented with both word and character representations but found no significant difference in results.

We evaluate our classifier the same way we describe in § 3.5. However, in the case of the percentage of correct scopes (PCS), we deem precision (and F_1 score alike) not to be informative given that our system almost never predicts a scope that is not present in the gold standard (hence, we did not find any false positives). Moreover, we did not find any case of false negatives where the scope is detected as not being present whereas there exists one in the gold standard. This leaves us only with partial matches, and recall becomes then a simple accuracy of the scope we exactly and fully match over the total number of scopes in the data. This is what we are going to report.

Baseline. In preliminary experiments, we noticed many sentences where negation scope was a single span delimited punctuation or sentence boundaries, as in (33), which we reported below for convenience.

(46) It helps activation, **not** inhibition of ibrf1 cells .

To assess how important this feature is, we implemented a simple baseline to replace the one we used in §3.4: in the case of the BioScope and the SFU, we mark the scope as all tokens to the right of the cue up until the first punctuation marker or sentence boundary; in the case of the CNeSp and ConanDoyleNeg, all tokens to the left and the right of the cue up until the first punctuation (or sentence boundaries). We

[8] Data from https://dumps.wikimedia.org/

[9] NLPIR: https://github.com/NLPIR-team/NLPIR

complement this punctuation-driven baseline with the sentence baseline ('S-Baseline' below), discussed in § 3.4.

3.9.3 Results

Results on the development set are shown in Table 3.5, including those on CONAN-DOYLENEG for comparison.[10] A boundary-based baseline is in most cases already a very strong system, especially in terms of recall. The relatively low precision is due to over-predicting the scope, which is trying to match sentence boundaries whilst the scope is a smaller span. As for the BIOSCOPE and the SFU corpora, the baseline yields the best performance in terms of PCS; this is due to the fact that both the CNESP and CONAN-DOYLE-NEG. annotates the subject (to the left of the cue in both English and Chinese) as part of the scope, whereas the remaining corpora do not.

In general, the BiLSTM system improves on joint prediction only in terms of PCS, mainly by predicting more continuous spans. This is also shown by a great reduction in the number of gaps in the predicted scope.

Finally, the system seems to perform poorly on CNESP-scientific and the BIO-SCOPE-full, which we believe is due to the small size of the corpus.

Results on the test set are shown in Table 3.6 and are compared to previous state-of-the-art-systems. In general, the performance of the boundary baseline mirrors the development set, where PCS outperforms both our and previously developed classifiers in three of the sub-corpora we considered. When this is not the case our system outperforms previous work, except in cases where training data is small (BIOSCOPE-FULL and CNESP-SCIENTIFIC.

Joint prediction again leads to a better PCS and scopes that do not contain gaps, whereas a vanilla biLSTM still yields better performance on token-based prediction.

3.9.4 Error analysis

The baseline results suggest that punctuation alone is a strong predictor of negation scope, so we further analyze this on the development set by dividing the negation instances into those whose scopes (in the human annotations) are precisely delimited by the innermost pair of punctuation markers containing the cue, and those which are not.

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

Data	System	Prec.	Rec.	F_1	PCS	gaps
Sherlock	Baseline	68.38	**85.82**	76.11	46.00	-
	BiLSTM	**93.65**	83.64	**88.36**(± 0.62)	**66.47**(± 0.71)	-
	+joint	91.80	79.62	85.22(± 0.3)	60.12(± 0.53)	-
SFU	Baseline	78.97	94.10	85.87	**79.00**	-
	S-Baseline	21.38	**100**	35.23	0.0	-
	BiLSTM	89.54	89.38	**89.45**(±0.17)	75.81(±0.4)	33
	+joint	**89.68**	86.75	88.16(±0.8)	77.33(±0.56)	**8**
BioScope Abstract	Baseline	86.24	82.21	84.18	**87.00**	-
	S-Baseline	29.49	**100**	45.55	0	-
	BiLSTM	**92.95**	89.56	**91.21**(±0.46)	71.37(±1.86)	35
	+joint	87.94	87.12	87.35(±1.52)	73.33(±1.77)	**0**
BioScope Full	Baseline	77.92	**86.12**	81.82	**73.00**	-
	S-Baseline	30.18	**100**	46.36	0	-
	BiLSTM	**87.67**	76.89	81.87(±2.73)	57.78(±5.38)	7
	+joint	78.88	71.82	74.48(±3.83)	42.42(±4.72)	5
BioScope Clinical	Baseline	94.24	93.92	95.08	**93.00**	-
	S-Baseline	55.95	**100**	71.75	0	-
	BiLSTM	94.40	96.62	**95.49**(±0.17)	89.17(±0.53)	10
	+joint	**94.94**	94.12	94.97(±0.14)	90.00(±0.49)	**0**
CNeSp Product	Baseline	71.71	96.10	82.14	40.00	-
	S-Baseline	12.11	**100**	21.61	0	-
	BiLSTM	**91.70**	88.47	**90.05**(±0.07)	63.78(±0.79)	33
	+joint	87.94	88.37	88.13(±0.69)	**66.20**(±0.17)	**0**
CNeSp Financial	Baseline	85.44	97.84	91.22	58.00	-
	S-Baseline	14.80	**100**	24.78	0	-
	BiLSTM	94.62	96.94	**95.76**(±0.39)	**75.76**(±0.81)	9
	+joint	**94.88**	93.50	94.17(±0.84)	60.61(±2.51)	**0**
CNeSp Scientific	Baseline	70.71	86.96	77.99	32.00	-
	S-Baseline	14.46	**100**	25.27	0	-
	BiLSTM	**88.55**	75.16	**81.06**(±4.48)	31.82(±3.21)	5
	+joint	85.11	68.07	74.58(±4.27)	**40.91**(±8.53)	**4**

Table 3.5: Results on the *dev* set for the English corpora (Sherlock, SFU & BioScope) and for Chinese corpora (CNeSp). ‘BiLSTM’ refers to the model described in §2 whereas ‘+joint’ refers to the model where a transition-based system is added on top. Standard deviation is reported for both the token-based and PCS F_1 measure.

Data	**System**	**Prec.**	**Rec.**	**F$_1$**	**PCS**	**gaps**
SFU	Baseline	80.26	91.29	85.42	79.00	-
	S-Baseline	22.98	**100**	37.37	0	-
	Cruz et al. (2016)*	85.56	82.64	84.07	58.69	-
	BiLSTM	90.76	88.76	**89.74**	77.71	8
	+joint	**90.92**	86.99	88.86	**79.05**	**0**
BioScope Abstract	Baseline	85.34	78.79	81.94	**86.00**	-
	S-Baseline	29.88	**100**	46.01	0	-
	Zou et al. (2013)*	-	-	-	76.90	-
	BiLSTM	**92.19**	90.77	**91.45**	70.44	30
	+joint	87.80	87.07	87.23	73.36	**0**
BioScope Full	Baseline	70.14	75.57	72.75	62.00	-
	S-Baseline	28.25	**100**	44.05	0	-
	Velldal et al. (2012)*	-	-	-	**70.21**	-
	BiLSTM	**84.47**	71.47	**77.39**	49.59	11
	+joint	75.04	66.02	69.14	42.15	18
BioScope Clinical	Baseline	97.18	97.39	97.29	**97.00**	-
	S-Baseline	57.03	**100**	72.63	0	-
	Velldal et al. (2012)*	-	-	-	90.74	-
	BiLSTM	97.71	98.04	**97.87**	94.74	8
	+joint	**98.78**	93.92	96.25	94.74	**1**
CNeSp Product	Baseline	70.22	97.57	81.67	42.00	-
	S-Baseline	11.81	**100**	21.14	0	-
	Zou et al. (2016)*	-	-	-	60.93	-
	BiLSTM	**91.11**	91.13	**91.12**	68.37	15
	+joint	89.50	91.37	90.39	**69.73**	**0**
CNeSp Financial	Baseline	84.80	97.52	90.51	54.00	-
	S-Baseline	15.38	**100**	27.67	0	-
	Zou et al. (2016)*	-	-	-	56.07	-
	BiLSTM	93.42	96.19	**94.78**	77.92	8
	+joint	**94.05**	95.00	94.50	**78.35**	**0**
CNeSp Scientific	Baseline	78.36	90.52	**84.00**	**64.00**	-
	S-Baseline	23.2	**100**	37.66	0	-
	Zou et al. (2016)*	-	62.16	-	62.16	-
	BiLSTM	**94.59**	58.71	72.21	31.82	3
	+joint	93.30	57.72	70.60	31.82	0

Table 3.6: Results for the English corpora (Sherlock, SFU & BioScope) and for Chinese corpora (CNeSp). * denotes results provided for context that are not directly comparable due to use 10-fold cross validation, which gives a small advantage in training data size.

Data	Punctuation	Other
Sherlock	66%	48%
SFU	80%	38%
BioScope Abstract	71%	40%
BioScope Full	44%	23%
BioScope Clinical	91%	48%
CNeSp Product	65%	40%
CNeSp Financial	72%	58%
CNeSp Scientific	20%	32%
Average	64%	41%

Table 3.7: PCS results on the development set, split into cases where punctuation exactly delimits negation scope in the gold annotation, and those where it does not. The results are averaged across all runs.

The results in Table 3.7. confirm a gap in accuracy between these two cases. The model correctly learns to associate surrounding punctuation with scope boundaries, but when this is not sufficient, it underpredicts, as in (47), or overpredicts, as in (48); prediction is again shown in curly brackets {}.

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

(48) {下次 是肯定 不 会 再 住 锦地 星座 了}
Next-time be surely not can again live Pingdi Xingzuo ASP
'Next time I won't live again in Pingdi Xingzuo for sure'

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

To understand why the system is so much better on punctuation-delimited scope, we examined the training data to see how frequent this pattern is found. The numbers in Table 3.8 suggest that our model may simply be learning that punctuation is highly indicative of scope boundaries, since this is empirically true in the data; the fact that the Sherlock and CNeSp-Scientific are the exception to this is in line with the observations above.

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

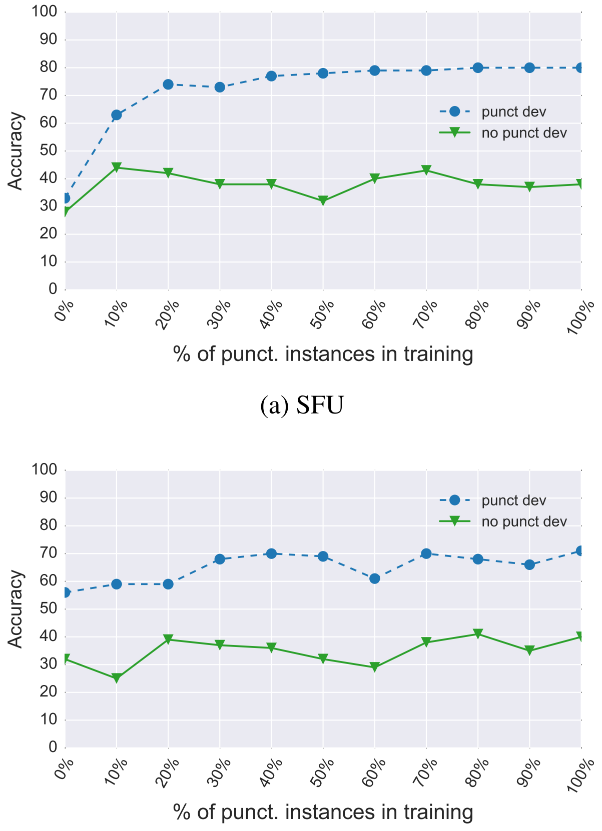
Table 3.8: Training instances by corpus, showing total count and percentages whose scope is predictable by punctuation boundaries only.

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

In order to see whether training imbalance was at play, we experimented with training by under-sampling from training examples that can be predicted by scope boundaries only. We report results on using incrementally bigger samples of the majority class. Figure 3.2 shows the results for the SFU and the BIOSCOPE-abstract corpora. There does indeed seem to be a slight effect where the classifier overfits to punctuation as delimiter of negation scope, but in general, classification of the other cases improves only slightly from under-sampling. This suggests that the absolute number of training instances for these cases is insufficient, rather than their ratio.

3.9.5 Re-annotation of negation scope

At this point it is worth asking: is negation scope detection easy because most of the instances in real data are easy? Or is it because the annotation guidelines made it easy? Or is it because of the domain of the data? To answer these questions we conducted a small experiment on SFU, BioScope-abstract and CNeSp-financial, each representing a different domain. For each, we randomly selected 100 sentences and annotated scope following the Sherlock guidelines. If the guidelines are indeed responsible for making scope detection easy, we should observe relatively fewer instances predictable



(a) SFU

(b) Bioscope-Abstract

Figure 3.2: PCS accuracy on the development set divided into instances where the punctuation and scope boundaries coincide (*punct.*) and instances where they do not (*no punct.*), when *punct.* instances are incrementally removed from the training data.

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

Table 3.9: Percentages of scope instances predictable (punct.) and not predictable (no punct.) by punctuation boundaries only on 100 randomly selected sentences annotated following the *Sherlock* guidelines for each of the three corpora considered.

by punctuation alone in these new annotations. If instead, easy instances still outnumber more difficult ones, we can conclude that detecting negation scope is less easy on Sherlock Holmes because of the domain of the data. Comparing the results in Table 3.9 with the one in Table 3.8, the Sherlock-style annotation produces more scopes that are not predictable by punctuation boundaries than those that are. We attribute this to the fact that by capturing elliptical constructions, the Sherlock guidelines require the annotation of complex, discontinuous scopes, as in (49).

(49) BioScope: second , t cells , which lack cd45 and **can not** signal via the tcr , supported higher levels of viral replication and gene expression .

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

In contrast with the original SFU and BioScope annotation, always annotating the subject produces negation scopes that are not bound by punctuation, since in both English and Chinese, subjects generally appear to the left of the cue and are less often delimited by any punctuation (50).

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

SFU-Sherlock:i ’m sure she felt rather uncomfortable having to ask us at all , but she thought it was strnge that we ’d **not** mentioned it.

3.10 Chapter conclusions

We conclude by answering the questions noted in the introduction of this chapter.

1. *Can we build a model that exploits the same set of features regardless of the language while performing as good or better than previous approaches?* **Yes, we can.** We showed that a bi-directional LSTM using only word and universal PoS tags embeddings available for a large number of languages achieve start-of-the-art performance on all corpora available annotated for negation scope when compared to previously developed non-neural systems.

2. *Is our model still state-of-the-art across-genre?* **No**. When training our system on Sherlock Holmes' stories and testing it on annotated sentences from Simple Wikipedia, we found performance to suffer across-genre, though performing still on-par with systems deploying richer feature representations. We also found our classifier to detect negation scope with higher accuracy around predicate negation than morphological negation.

3. *Does the same approach work equally well across all available corpora, one of which in Chinese?* **Yes**; however, for most corpora, this can be attributed to a single fact: **negation scope is annotated as a single span of text delimited by punctuation**. For negation scopes not of this form, detection accuracy is low and under-sampling the easy training examples does not substantially improve accuracy. We demonstrate that this is partly an artifact of annotation guidelines.

Chapter 4

NegPar: a parallel corpus annotated for negation

This chapter is based on the following peer-reviewed paper:

Qianchu Liu, Federico Fancellu and Bonnie Webber (2017), *NegPar: a parallel corpus annotated for negation*, to appear in the Proceedings of LREC2018.

4.1 Introduction

So far, we have considered the task of negation scope detection in the presence of annotated data. However, creating a corpus annotated for negation is a time-consuming task, as it is adapting existing guidelines for English to other languages to ensure high inter-annotator agreement (e.g. Altuna et al. (2017)). Perhaps for this reason, there exists only one annotated corpus for Chinese, the CNESP.

It is then worth asking: *to what extent do we need annotated data in a target language to detect negation scope*? This and the following chapter will try and answer this question by looking at model transfer, where we train a model in a 'source' language where annotations are available, English, and then test it in a 'target' language, where these are not available. In this thesis we use Chinese as our target language.

We start, perhaps in a somewhat reversed order, with the problem of the data our cross-lingual model should be trained and tested on. One prerequisite is that the data should be annotated the same way in both the source and the target. As we saw in §2.5, none of the available corpora in English and Chinese satisfy this requirement, with the annotation of the CNESP, being based on but not fully adhering the BIOSCOPE guidelines.

To fill this gap, we present NEGPAR, the first parallel corpus annotated for negation. The corpus has been created by leveraging the CONANDOYLENEG corpus (Morante and Daelemans, 2012), where the annotation of cue, scope and event are extended onto a sentence-aligned Chinese translation. We chose Chinese to start by looking at a pair of languages where negation is expressed similarly. As for the data, we chose CONANDOYLENEG because human translations of the relevant Sherlock Holmes' stories are freely available.

The annotations consisted of two different stages.

First, we have reannotated the original CONANDOYLENEG corpus to cover those phenomena that the original annotations did not consider or we thought should be handled differently.

Second, we annotated the Chinese side. To ease the annotation task, we experimented with using word alignment based annotation projection to transfer pre-existing annotations on English sentences onto their Chinese translations. We then manually correct the projection results. Automatic projection allowed us to both investigate and quantify the linguistic differences in representing negation between the two languages. We are then asking: *in the ideal scenario where annotated parallel data in both English and Chinese is available, how easy is to detect negation across languages?.*

We evaluate the results of the projection by measuring precision, recall and F_1 measure over the tokens projected. Results have shown that projecting annotations across languages via word-alignment does not yield good results for any of the subcomponents of negation. In particular, the low recall has revealed that Chinese exhibits a large number of negative instances that are translated as positive constructions in English. Through an error analysis, we present some examples of these translation divergences between English and Chinese, as well as example of errors caused by alignment errors.

4.2 The Annotation Task

4.2.1 The corpus

Our annotated parallel corpus aligns the four stories annotated in CONANDOLYENEG ('The Hound of the Baskervilles', 'The Adventure of Wisteria Lodge', 'The Adventure of the Cardboard Box' and 'The Adventure of the Red Circle') with their Chinese translations by Mengyuan Lin.[1] The format of the annotated Chinese side is the same CoNLL format as used in CONANDOYLENEG, where each negation instance in a sentence appears in a set of three columns, for cue, event and scope respectively.

The annotation was carried out by a native Mandarin speaker with a background in linguistics and the guidelines were created by the annotator and the author of the present thesis. The latter is also responsible for the reannotation of the English side.

The Chinese side of NEGPAR was pre-processed twice, first to segment the sentences into words and second to correctly align the sentences to their English source.

To segment sentences into words we use the Stanford Segmenter (Chang et al., 2008); however automatic segmentation is not always consistent or correct. For this reason we also experimented with projecting and annotating at a character level, where English words are mapped into Chinese characters.

To create a parallel corpus we used the sentence aligner Hunalign (Varga et al., 2007) which was trained using the English side of the corpus and a supplementary English-Chinese bilingual dictionary CEDICT[2]. To increase the accuracy of the sentence aligner, we split each story in the corpus into chapters, which correspond across the two languages. Finally, to ensure the minimal effect from sentence alignment errors on the gold annotations and the projection results, all sentence alignment errors that involve negation were manually corrected.

To further improve alignments, we moved all right-attached inter-quotation attribution phrases in Chinese to be left-attached, which is the norm in English. This is exemplified in (51) where the attribution phrase ‘他说道’ (‘he said’) is attached to the preceding direct quotation (square brackets [] stands for sentence boundaries):

(51) ‘就在这里好了。’ 他说道,‘右侧的这些山石是色妙的屏障。’
[‘就在这里好了。’ 他说道。][‘右侧的这些山石是色妙的屏障。’]
[‘This will do’, said he.] [‘These rocks upon the right make an admirable screen’]

A quantitative comparison of the English and Chines side of NEGPAR is presented in Table 4.1. Overall, there are more negation instances in the Chinese translation than in English as the percentage of negated sentences, the number of cues, scopes and events in Chinese are consistently higher.

[1] The Chinese translation can be found at http://www.kanunu8.com/book3/8105/.

[2] `https://www.mdbg.net/chinese/dictionary?page=cedict`

	# sents	# neg. sents	# cues	# events	#scopes
English	5520	1227(22.22%)	1421	911	1304
Chinese	5005	1442(28.81%)	1782	1168	1762

Table 4.1: Comparison between the English and the Chinese side of NegPar in terms of total number of sentences (# sents.), negated sentences (# neg.sents.), number of cues (#cues), events (#events) and negated scope spans (#scopes).

4.2.2 Rereading the annotations on the English side

In Ch. 2 we mentioned some of the inconsistencies in the original ConanDoyleNeg annotations. Before projecting these onto Chinese, we thought it would be desirable to reread the guidelines of Morante et al. (2011) and annotate some of the phenomena related to negation scope differently.

We divide this discussion in two parts, the first addressing the phenomena we thought deserved a different treatment and the second addressing those that were not considered at all by the original guidelines.

4.2.2.1 Alternative annotation choices

Morphological negation. Although the guidelines address morphological negation (as in '**un**happy', '**im**patient', etc.), this does not involve scope, so although keeping the negation suffixes annotated, we do not consider them any further. However, as a matter of consistency, we reconsider the status of those adverbs containing a negation affix. This is because, whereas the original English guidelines state that 'If the negated affix is attached to an adverb that is a complement of a verb, the negation scopes over the entire clause'(Morante et al., 2011, p.21), we found cases in the corpus where it is just the adverb to be included in the scope, as shown when comparing the examples in (52).

(52) a. [...] tossing rest**lessly** from side to side

b. [...] glaring help**lessly** at the frightful thing which was hunting him down

In general we felt that a wide scope is not ideal in the case of adverbs since the event actually takes place but it is just the manner the event takes place that is negated.

For this reason, we just **annotate the adverb as being in the scope of negation**; such cases were therefore reannotated in the same way as (52.a).

Except/save/no...but. Exceptions are also another phenomenon we felt the need to reannotate. Morante et al. (2011) state that 'often [these items] function as neutralizers of the polarity of the statement [...] where they occur', which is why they are considered as part of the scope, along with the material they introduce. This is exemplified in (53):

(53) [...] Marx knew **nothing** of his customer save that he was a good payer

This however does not reflect the fact that 'save' excludes from the set of things negated in the matrix clause, which should interpreted as positive. (53) can in fact be paraphrased as '*It is not the case that* (Marx knew nothing about his customer). He knew that he was a good payer'.

This use of 'save' contrasts from its use as negation cue, as shown in (54).

(54) Mr. Sherlock Holmes, who was usually very late in the mornings, **save** upon those not infrequent occasions when he was up all night, was seated at the breakfast table.

Here, 'save' is used to neutralize positive polarity and to highlight the set of instances where an event did not take place; (54) implies in fact that 'he was usually very late but he wasn't late on those not infrequent occasions when he was up all night'.

For these reasons, when annotating we distinguish two types of exceptions. **The 'exception to negative', usually positive as shown in (53), where we exclude the exception from the scope of negation; and the 'exception to positive', with a negative meaning as shown in (54), where we only include the 'except' phrase in the scope of negation, as the original guidelines already do**.

Subordinate clauses. The original guidelines state that 'when a verb is negated the whole clause is in the scope of negation', including any subordinate clause. This means that all subordinates is included in the scope; let us take the following example, alongside its original annotations.

(55) After what we heard, I do**n't** feel as if I could give the man up

We have here two events in temporal order, 'hear' and 'give up' (here in a neg. raising construction) and where only the latter is negated. Including both events in the scope of negation might lead to the interpretation that 'nothing was heard' ('*It is not the case that* I feel as if I could give the man up after what I have heard, because nothing was heard'), which is not in line with the meaning of the sentence in (59). The wide scope interpretation is also in contrast with the event-centric idea of negation scope where it is such "that it allows us to determine which events are negated in the sentence"(Morante et al., 2011, p.13). This is even more evident in the examples below, annotated according to the original guidelines:

(56) a. He did **not** go because all roads were closed

b. He did **not** go because he wanted to but because I forced him.

In the first case, the reason has no bearing on the 'go' event being negated, whereas it is only the reason to be negated in the second clause, with the event 'to go' still happening. The original guidelines do not make this distinction and end up annotating the matrix clause even when it is only the subordinate to be in the scope.

Given these considerations, we opt here to **exclude all subordinate clauses from the scope of negation unless directly negated, in which case we do not include the matrix clause in the scope but only the subordinate conjunction**; this is shown in the reannotation of (56.a) and (56.b) respectively.

(57) a He did **not** go because all roads were closed

b He did **not** go because he wanted to but because I forced him.

To support our reannotation, previous work has also shown that humans tends to interpret these constructions as having small scope, i.e. scoping only on the matrix clause (Khemlani et al., 2012).

4.2.2.2 Annotating previously excluded phenomena

Neg raising. Neg raising – the phenomenon that a negation in the matrix clause is interpreted in negating the complement clause, is not covered by the annotation guidelines. Neg raising is encountered with verbs expressing the speaker's opinion, such as 'think', 'believe', 'want', 'seem', etc.. In cases like (58), the annotations consider the entire sentence under the scope of negation; however, it is not the thinking that should be negated but the object of the thought.

(58) I do **not** think it is likely = I think that it is not likely.

In cases of neg raising we annotate as part of the scope the subordinate only. (58) is therefore reannotated as (59)

(59) I do **not** think it is likely

Quantifiers. The interaction between quantification and negation scope at a string level is not considered at all in the original ConanDoyleNeg guidelines.

Cases where ‘not’ directly precedes lexical items like ‘all’ and ‘every’ are correctly annotated, as demonstrated in the following example.

(60) Money is **not** everything.
(= *It is not the case that* money is everything).

However, let us consider the following example, annotated according to the original guidelines.

(61) The fellow might have had other reasons for thinking that all was **not** well

The original guidelines paraphrased the construction under the scope as ‘It was not the case that the fellow was thinking that all was well’; however, from a logical perspective, the universal quantifier should scope over negation and not viceversa, given that $\forall thing(x)\neg \rightarrow \exists s.well(s) \wedge Topic(e,x)$

In cases like (62), we **exclude the lexical item representing universal quantification** to yield the following annotation.

(62) The fellow might have had other reasons for thinking all was **not** well

Modals. The interaction between the scope of negation and modality is another phenomenon the guidelines do not mention. Some cases, as the one shown in (63), are correctly handled, where negation correctly scopes over the modal.

(63) You need **not** to fear to speak the truth. = *It is not the case that* you need to fear to speak the truth.

We however found two cases of deontic modality where the annotations fail to capture this interaction as shown below.

(64) You certainly must **not** go alone $\neq$ *It is not the case that* you certainly must go alone.

Having negation taking scope over ‘must’ leads to the incorrect interpretation where the person could go alone.

In cases like (64), we adopt a strategy similar to the one used for quantifiers and **exclude the lexical item representing modality from the scope**. This leads the example in (64) to be annotated as follows:

You certainly must **not** go alone

4.2.3 Annotating negation in Chinese

We include here a brief summary of the annotation guidelines in Chinese, where we report annotation examples subdivided into the three components we considered: cue, scope and event. The annotation guidelines in full can be found in Appendix B.

4.2.3.1 Cue

We annotated a total of 45 negation cue types in Chinese including adverbs, auxiliary verbs and prefixes. Amongst these we found 10 *core* negation cues (the same defined in §2.3; for a definition of *core* cues we refer the reader to the same section). The most common cue in Chinese is the adverb 不 (roughly equivalent to ‘not’) which is used both as a stand-alone word and as an affix (e.g. 不贵,literally ‘inexpensive’).

Infix cue in verb-complement constructions. In Chinese the negation cue 不 can appear as an infix in verb-complement construction. These complements usually indicate the result, the direction or the manner of an action expressed by the main verb, as well as expressing potential forms (roughly equivalent to the English ‘cannot’). In this regard, Li and Thompson (1989) give an interpretation of infixal negation in Chinese in that the result introduced in the complement is ‘unachievable’. (65) examplifies this construction alongside its annotations, where only the infixed cue and neither the verb nor the complement are considered cues.

(65) 他 说 得 清楚
He speaks ADV clear
‘He speaks clearly.’
他 说 不 清楚
He speaks not clear
‘He does **not** speak clearly.’

Non-functional negation cue. We do *not* annotate any non-functional negation, i.e. expression that include a negation cue but have positive meaning.

Certain fixed expressions belong to this category; as shown in (66), the expression ‘can’t help’ and the Chinese counterpart ‘不得不/不能不’ , despite including a negation marker have a positive meaning (i.e. the action they specify has or will take place).

(66) 我 不得不 讨厌 他
I not-should-not hate he
I couldn’t help but hating him

Certain idioms consisting of four characters, may be false negatives as well. This is true in particular when the first and the third characters are negation markers, like 无往不利, literally ‘there are no places where victory is not achieved’, with the positive meaning of ‘always successful’.

Similar to English, another problem arising from non-functional negation is identifying negation affixes that do not introduce negation. For instance, in the word ‘disgrace’ the affix ‘dis-’ is *not* considered a negation marker because the meaning of the whole word is not ‘lack of grace’; on the other hand, ‘impatient’ is opposed to ‘patient’.

This problem is mostly related and can be solved through *semantic transparency*; for these reason in words like 无聊‘boring’, literally ‘no chatting’, we do not annotate the cue 无 as cue.

Another criterion that we used is *obsolescence*: if the meaning of a word modified by a negation affix is now obsolete, we do not annotate the affix as cue.

Finally, we exclude negation cues used in rhetorical questions, which often take the form of ‘modal+cue+modal’, roughly equivalent to the English ‘shall we...?’

(67) 咱们 要 不 要 向 后 退
we want not want towards back retreat
‘Shall we move further back?’

Discontinuous cues. Certain cues in Chinese are discontinuous, similar to the English construction ‘neither...nor’, the equivalent being the construction ‘既不...也不’. It is worth mentioning however that omission is a feature of these constructions which can be reduced further to ‘不...不’, therefore preserving only the core cues.

(68) 对 他 **(既)不** 应该 可怜 , **(也)不** 应该 原谅
towards him not should pity , not should excuse
[...] for whom there was neither pity nor excuse

4.2.3.2 Scope

Sentential negation. If negation is sentential, i.e. the predicate of a simple clause is negated by the cues ‘不’ and ‘没(有)’, we annotate the entire clause under the scope of negation. In the case of two or more coordinated clauses where only one is negated, we annotated as inside the scope only the negated clause whereas the others fall outside of it. If there is any material omitted from the negated clause that is retrievable from other parts of the sentence, this is annotated as well. (69) exemplifies the annotation of coordination.

(69) 我 把 他 弃 而 不 顾 了
I BA him abandon and not care ASP
‘I abandoned and did not care about him’

Subordination. If it is the event in the subordinate clause to be negated, this is included in the scope of negation unlike the matrix clause which is excluded. On the other hand, if the event negated is in the matrix clause, subordinates are usually excluded from the scope of negation.

However, Chinese allows for it-cleft constructions like the one in (70), where only the subordinate clause, which appears before the event of the main, is in the scope of negation.

(70) 您 不 会 因为 知道 了 这 一点 而 感到 高兴
you not can because know ASP this point then feel happy
It is not because you know this that you feel happy.

Relative clauses. If negation appears in a relative clause. we annotate only this in the scope of negation but not the head noun that it modifies. Unlike English, where the clause follows the head, Chinese displays the opposite order, with the particle ‘的’ in between. This is exemplified in (71), where ‘的’ separates the relative clause ‘不爱出风头‘, ‘to like to show off’, and the head ‘人’, ‘person’.

(71) 他 是 个 不 爱 出风头 的 人
he is CL not like show-off DE person
‘He is a person who does not like to show off.’

Nominal and adjectival predicates. When negation directly denies a state which is also the main predicate of a clause, the scope is over the entire clause. Whereas in English, these constructions are formed by the copula followed by an adjective (”He is impatient”), Chinese does not require a copula. This is shown in (72).

(72) 这样 不 公正
This way not fair
‘This is unfair.’

In relation to these constructions, one important difference between Chinese and English is the status of affixal negation. If in English affixal negation creates contraries and not contradictions, hence not forming a scope, in Chinese an adjective and its negated counterpart cannot be false at the same time, therefore abiding by the ‘Law of the Excluded Middle’. This is exemplified as follows:

(73) I am neither patient nor impatient
* 我 既不 耐心 也不 不耐心
I neither patient nor impatient

As for nominal predicates, where a noun phrase follows a copula (similar to the English ‘He is not a patient man’), we also annotate the entire clause in the scope, as shown in (74)

(74) 他 不 是一 个 耐心 的 人
He not is one CL patient DE man
‘He is not a patient man’

Sentence final particles. Chinese is characterized by sentence-final mood particles that express the attitude or mood of the speaker towards the whole sentence. Given that these particles are not affected by the presence of a negation cue, they are *not* included in the scope of negation; this decision is also supported by theoretical work that define these particles as complementisers out of the IP (Paul, 2014). For example in (75), sentence final ‘呀’ only emphasizes the content of the clause it ends but is not itself affected by negation.

(75) 不 要 等 他过 了 山 呀!
not need wait he passed ASP mountain MOOD
‘There is no need to wait until he has past the mountain!’

Comparative constructions. In Chinese, comparison is expressed in most cases through the co-verb ‘比’, which takes as subject and object the two things compared, followed by the dimension they are compared along. This is the case in (76), where the subject and the object are compared for their age; in cases like this, we annotate as scope the entire clause.

(76) 约翰森 先生 年纪 不 比 你 高
Johnson Mr. age not compare you old
‘Mr. Johnson is not older than you.’

However, negation can also exclude this dimension as shown in (77). We distinguish this from (76), by excluding from the scope in the object of the comparison, as shown in the example below.

(77) 我 的 觉 睡的 比 平常 还要 不 踏实
I of sleep sleep compare normal even not easy
‘I haven’t slept as deeply as I usually do’

4.2.3.3 Event

We annotate an event as negated if it is factual; the term ‘factuality’ includes here both states and nominal elements as well. What the annotation considers as event is a minimal unit in a negated phrase, usually corresponding to its head. An example of annotation of a verbal predicate negated event is shown in (78), where the event is presented inside a box (we omit the scope just for presentational purposes). Although one could consider 吃羊肉, ‘eat mutton’, as the entire event, the event is just its minimal unit, that is, the head verb 吃 ‘to eat’

(78) 我 不 吃 羊肉
I not eat mutton
‘I do not eat mutton.’

Existential and copulative constructions. In existential constructions, we do *not* mark the verb 有, ‘there is/are’, as an event; instead, we mark as the event the head of the nominal phrase following the existential construction as shown in (79).

(79) 这里 没 有 人
here not there-is people
‘There is nobody here’

As shown in (79), the existential construction in Chinese also encode universal quantification (i.e. ‘nobody’). When universal quantification applies to the subject of the clause, we annotate as event the head of the verbal predicate. In (80), we therefore mark as the event ‘动’, ‘to move’ but not the aspect marker ‘在’, marking a continuous action.

(80) 没 有 人 在 $\boxed{\text{动}}$
not there-is person ASP move
‘Nobody is moving’

Modality. Given that we annotate factuality, we do *not* annotate as events those verbs in the scope of certain modals, in particular where the speaker is uncertain about the happening of an event. In English, this excludes most cases of epistemic modality (i.e. verbs introduced by auxiliars such as ‘should’, ‘would’, etc.). Similarly in Chinese, we do *not* annotate negated events in the scope of modals except for modality expressing the subject internal ability. This is the case of the modal ‘能’ which is annotated only when expressing participant internal abilities (81.1) but not when expressing conjecture about a non-factual event (81.2).

(81) 1. 我 不 $\boxed{\text{能}}$ 打 篮球
i not can play volleyball
‘I cannot play volleyball’

2. 我 不 能 忍住 这 种 情况
i not can bear this CL situation
‘I couldn’t bear this kind of situation

Supposition or presumption. In order to determine whether to annotate something as an event, we also examine the semantics of the verb that directly follows the cue. If the verb suggests that the speaker is certain about the content of the predicate, we treat the head of the predicate as factual and annotate the negated event in the clause. On the other hand, if the verb suggests that the predicate is only supposed or presumed by the speaker, we do annotate the head of the predicate as event. This contrast is exemplified in (82.1) and (82.2), through the verbs ‘知道’ , ‘to know’, where we annotate the event, and ‘相信’ , ‘to believe’, where we do not.

(82) 1. 我 知道 您 决不 愿意 $\boxed{\text{做}}$ 一 个 妨碍 别人 的 人
i know you not want do one CL hinder others DE person
‘I know that you do not wish to be a spoilsport’

2. 我 相信 您 决不 愿意 做 一 个 妨碍 别人 的 人
i believe you not want do one CL hinder others DE person
‘I believe that you do not wish to be a spoilsport’

4.3 Annotation projection

4.3.1 Background

Previous work have investigate whether in presence of parallel data or word-alignment information it is possible to transfer annotations from a resource-rich language (usually English) to one where such resources are scarce. As an alternative to direct model transfer, this process of *annotation projection* has been widely used for different types of annotations in NLP. These include amongst other:

- *Syntactic parsing*, where word-alignment information is used to project syntactic information from English to a target language where a parser is then trained (Hwa et al., 2005).

- *Semantic role labeling*, where both word-alignment has been used to map semantic roles across languages. In this line of research, work have experiment with constraining projection using syntactic information, in the form aligned constituents (Padó and Lapata, 2009).

- *Analysis of translation divergences*, where word-alignment based transfer of dependencies has been used to measure the degree of similarity between translation pairs (Hwa et al., 2002).

- *Word senses*. Diab and Resnik (2002) and Bentivogli and Pianta (2005) used word-alignment information to transfer word senses across language, under the assumption that words identified as being translation of each others tend to invoke the same concept

- *Coreference*. Postolache et al. (2006) experimenting with transferring coreference chains from English to Romanian via word-alignment information extracted from parallel texts.

4.3.2 Methodology

The goal of annotation projection is to investigate whether we can ease the burden of annotating from scratch in the presence of parallel text.

Annotations are projected using word alignment information computed using the IBM model 2, as implemented in the fast_align toolkit (Dyer et al., 2013).

	Word-level			Character-level					
	precision	recall	F1	precision	recall	F1	en	zh	proj.
cue	0.39	0.47	0.43	0.49	0.42	**0.45**	175	230	169(73%)
event	0.37	0.32	**0.34**	0.40	0.27	0.32	123	153	103(67%)
scope	0.64	0.48	**0.55**	0.64	0.44	0.50	170	226	168(73%)

Table 4.2: Performance of the annotation projection on the *dev* set both at word and character level for cue, event and scope respectively. We also report the number of gold scope spans in both English and Chinese as well as the number of spans projected from the former to the latter.

The aligned sentence pairs in NegPar are joined with the English-Chinese UN parallel corpus (Ziemski et al., 2016) to form the training data for the alignment model. The Chinese side of the corpus was tokenized using the Stanford Word Segmenter (Chang et al., 2008). We used a symmetrical two-way alignment as the basis for projection. In our work we experimented with two types of alignement model: an English word to Chinese word (word-level projection) and an English word to Chinese characters (character-level projection).

Following the work of Padó and Lapata (2009), one could argue that simple word-alignment projection could be constrained by checking whether the aligned spans share the same syntactic constituent. However, this proves difficult in our task for two reasons: 1) cue and event are usually made up of a single character or word and therefore syntactic information would be of no help; 2) scope is often a discontinuous span and syntactic constrains would therefore force to consider the entire constituent with no gap as in the scope of negation.

To ensure a fair comparison, we evaluated both levels of projection at character-level. We report precision, recall, F1 measure and number of gold and projected spans for cue, event and scope independently, as they were projected as such.

4.3.3 Results

The results for cue, scope and event projection on the development set are summarized in Table 4.2.

Considering F1 alone, we found word-level projection to yield better results for event and scope but not for cue projection. This can be explained by the fact that cues often span subword units (as in the case of morphological negation) and word-level projection might end up over-predicting cues in Chinese (hence the relatively lower precision). However, as for easing the annotation process, a relatively lower recall means that more work is required to find elements that the projection has missed, which is more burdensome and time-consuming than filtering out over-predicted elements.

In general, even in the presence of parallel data, detecting negation using annotation projection does not lead to good results. The number of projected spans vs. gold spans in Chinese and English suggests that this is in part due to differences in how negation is translated.

4.3.4 Error Analysis

We carry out an error analysis to delve deeper into errors made during projection. In doing so, we consider character-based projection for the cue and word-based projection for event and scope spans.

Cue. We first break down the performance of cue projection according to different Chinese cues. We found performance to vary across different cues as shown in Fig. 4.1. Compared to precision, recall is lower for two-character cues where projection often seems to miss either of the characters; this is the case of both 没有 and 并不.

The low performance of 无 and 未 might be caused by these two cues being common components in Chinese idioms that are rendered as positive constructions in English, as shown in (84). We found eight such cases.

We then analyzed projection errors based on English cues as shown at the bottom of Fig 4.1. In general, when the English cues are correctly projected onto Chinese, the projection also annotates additional surrounding characters in Chinese; therefore the recall is higher than precision. It is especially the case for the negation pronoun 'nothing' which maps to the negative polarity item (NPI) 什么都/一点也/一点/一, 'any/anything'. This accounts for 5 errors, one of which we report below (*proj.* is the projected annotation, *gold zh.* the Chinese gold standard and *gold en.* the English gold standard).

(83) *proj.*: 从他那里**什么 都** 得不到
gold-zh: 从 他 那里 什么 都 得 **不** 到
from him there anything DOU get not POT
*gold-en:*can get **nothing** from him

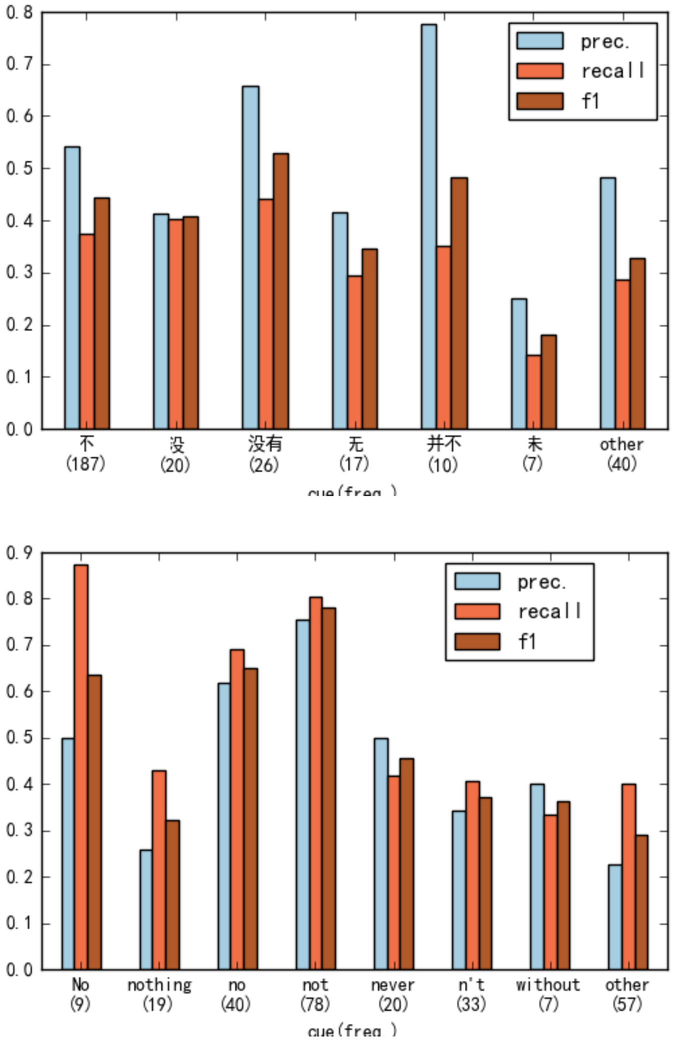


Figure 4.1: Performance of annotation projection plotted against Chinese (top) and English (bottom) cues

Finally, we wanted to analyze those cases where projection fails due to the fact that negation is present in Chinese but not in English. we found this to happen for two different reasons:

1. A negation instance in Chinese is paraphrased in positive terms in English; this often concerns just a pair of contrary adjectives or adverbs as shown in (84), but also extends to entire clauses as shown in (85)

(84) *gold zh.*: 他 安然 无 恙
he safe not sick
gold en.: He is safe and sound.

(85) *gold zh.*: 惊慌 的 脸上 没 有 一点 血色
panicked DE face not there-is one-bit blood-color
gold en.: Every tinge of colour struck from his astonished face

2. Some lexical items in English can be interpreted as inherently expressing negation and thus can be translated as cues in Chinese, but they are not annotated as cues on the English side. This is the case in (86); along with 'hardly', we found other expressions such as 'rather than', 'absence', 'out of question' and 'refused' that are translated into negation cues in Chinese.

(86) *gold zh.* 这 件 事 的 前前后后 不 可能 是 为了 [...]
This CL thing of everything not can be for [...]
gold en.: The whole proceeding could hardly be for [...]

Event. Out of the 153 gold events we found that only 13.3% were correctly projected from English, with 38.8% of the cases where the projection does not detect an event at all. These false negatives are caused by the fact that negation is present in English but not in Chinese (same as the case (84)∼(86)) but in some cases are just due to English words aligning to a null token.

On the other hand, in 16% of the cases we observed that the English event is projected onto a completely different span of the sentence. Some of these cases are however not due to alignment errors but because the Chinese side uses a different constructions with a different type of event from English. For instance, in (87), the English guidelines annotate as event the nominal predicate 'colour' where this is translated in Chinese as a verb 说上, 'say'.

(87) *gold zh.*: 那 张 脸 既 不 黑 [...] 说 不
上 是 什么 颜色
That CL face either not black [...] say not
up be what colour
gold en.: It was **n't** black [...] **nor** any **colour**

Finally, in 13.3% of the cases, the projection only partially matches the gold annotation for an event. In 13 of these 25 cases we found that the projection includes the negation cue inside the event. This is often the case, such as (88), where a word containing morphological negation in English is projected onto both the cue and the event in Chinese.

(88) projected: ’[...] 我们还弄不清楚 的罪行’
gold-zh: [...] 我们 还 弄 不 清楚 的 罪行
[...] we still manage not clear DE crime
gold-en: ’They were all confederates in the same **un**known crime .’

Scope. Out of the 226 instances of negation scope we found that only 3 (0.01%) were fully and correctly projected, with 39% of the cases where the projection returns nothing, once again due to negation instances in English being rendered as positive statements in Chinese.

We found only 5 cases (0.02%) where the scope in English is projected to a completely different span from the gold span in the Chinese sentence. The majority of the errors (145/226 – 77%) concern partial overlap, where the projection covers the gold scope only in part. A closer analysis shows that the projection tends to often miss the NPI 什么 (12 cases). 什么 corresponds to the English ‘any’ when in the scope of negation (otherwise its literal meaning is the interrogative pronoun ‘what’). In all the cases where projection fails to include this element in the scope, English uses the determiner ‘no’ or pronoun ‘nothing’ instead of an overt NPI such as ’any’; therefore什么 is mapped to the negation cue rather than being marked as scope element; this is exemplified in (89).

(89) projected: 这里面**没有什么**
gold-zh: 这里面 **没** 有 什么
Here not there-is anything
gold-en: **Nothing** in all this

Finally, in order for future work to compare with this baseline, Table 4.3 and 4.4 report the performance of annotation projection on the test set as well.

	precision	recall	F1
cue	0.372	0.428	0.398
scope	0.574	0.381	**0.458**
event	0.299	0.209	**0.246**

Table 4.3: Results from word-level projection of negation on the test data

	precision	recall	F1
cue	0.478	0.382	**0.425**
scope	0.583	0.312	0.406
event	0.338	0.180	0.235

Table 4.4: Results from character-level projection of negation on the test data

4.4 Chapter Conclusions

We have introduced NEGPAR, the first English-Chinese parallel corpus annotated for negation. NEGPAR is based on a pre-existing annotated corpus for English, CONANDOYLENEG, whose annotations we have extended onto its Chinese translations. Our contribution were as follows:

1. We reconsidered some phenomena related to negation scope that we though were not sufficiently brought to light or were not taken into account at all in the original CONANDOYLENEG corpus and reannotated those.

2. We provide an annotated corpus for Chinese, alongside annotation guidelines.

3. We experimented with automatic annotation projection via word-alignment to assess whether we can ease the annotation task, as well as to better understand the differences in signaling negation across these two languages. Results showed that automatic projection is of little help to the annotation process; in particular the relatively lower recall highlights how negation instances in the Chinese are missed because they correspond to a positive construction in English. Finally through an error analysis we gave some examples of these translation divergences as well as examples of alignment errors.

Chapter 5

Detecting Negation Scope Across Languages Via Universal Dependencies

5.1 Introduction

Using the NegPar corpus, it is now possible to address the question of whether a model for negation scope detection can be trained in a source language and used in a target one where annotated data is not available. To bridge the gap between source and target language we use two intermediate representations: Universal Dependencies (De Marneffe et al., 2014, UD below), a syntactic annotation framework consistent across languages, and cross-lingual word embeddings. A model can be trained to detect negation scope on top of UD parses in a language and tested on the UD parses in another. As shown in Figure 5.1, this is possible because the relation between negation scope and its surrounding dependency context is consistent across language; in this case, our model should learn that when a cue directly negates a parent event under conjunction (”drink”/ ”喝”) the scope spans the conjunction only, and that long-range material inside the scope is likely to be present (the subject ”I”/”我“).

We experiment here with two different neural models that accept trees or graphs as input (for convenience we will call these *structured models*): an extension of a recursive child-sum TreeLSTMs (Tai et al., 2015), here referred to as Bidirectional Dependency LSTM (BiDLSTM in short), and Graph Convolutional Networks (GCN Marcheggiani and Titov, 2017). We use the BiLSTM we developed in Chapter 3 as baseline; after observing how punctuation boundaries are used as the main feature for scope detection,

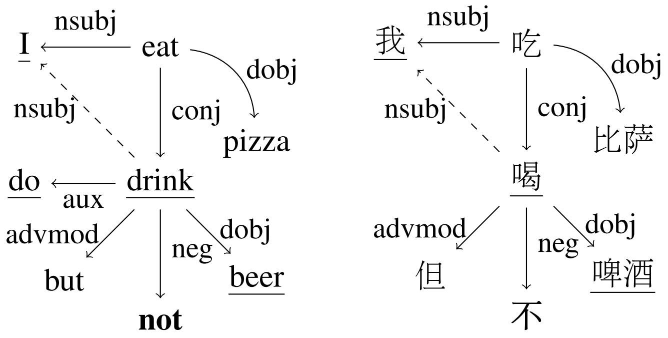


Figure 5.1: Dependency parse (using UD v1) for the sentence ‘I eat pizza but do not drink beer’ and its Chinese translation. The dashed line represent an enhancement to the parse available in the Stanford Enhanced++ representation (Schuster and Manning, 2016)

we also implement a variant of the model where punctuation was removed.

We conduct two series of experiments. We first assess the validity of the models in a monolingual setting, where we train and test in the language.

Results show that in a monolingual setting the BiLSTM are still state of the art in both English and Chinese; BiDLSTM on the other hand always outperform GCN. However, both structured models are not driven by punctuation or sentence boundaries in making predictions, whilst the BiLSTM still is.

We then test the model across language, by training in English and testing in Chinese. In comparison, we expect the performance of these recurrent classifiers to be worse by a large margin in the cross-lingual task, given its reliance on language specific word order information. Result show that it is indeed possible to build a model transferable across language, with performance close by not as good as a monolingual oracle. However our main finding is that **using pre-trained cross-lingual word embeddings only help little with the task of detecting negation scope across languages, proving the usefulness of UD for cross-lingual semantic tasks; instead, the model relies mostly on information coming from the dependency structure to guide prediction**. This is because negation scope is concerned mostly with structure and less with lexical semantic similarity. In the BiLSTM, where explicit parse information is not available, the model uses punctuation and sentence boundaries as proxies for structure to predict the scope.

Finally, error analysis shows that structured models perform better when the cue is in the same dependency substructure as its scope but still fail to predict some phenomena related to negation scope, especially when lexical information is required.

5.2 Background

5.2.1 An overview of model transfer approaches

In the absence of annotated data in a target language, many work have experimented with model transfer, where a model is trained in a language and tested in another. Despite the common scenario of having to deal with data sparsity in a language other than English, model transfer differs from annotation projection in that the latter explores way of transferring only annotations but not the model itself across languages.

In general most of the work have looked at model transfer for dependency parsing (e.g. Tiedemann, 2015; Ammar et al., 2016, amongst the most recent work), given that annotated data is available for a large number of languages other than English. The pioneering work by Zeman and Resnik (2008) perfectly exemplifies the idea of model transfer for a syntactic parser, where a (non universal) dependency parser is trained in a language and tested on a related one, where discrepancies between labels are fixed using hand-built heuristics.

In place of heuristics to deal with discrepancies in the annotations between the train and and the test language, many work have opted for language-independent representations or features. Cohen et al. (2011) for instance trained an unlexicalized parser on multiple source languages and applied to several target languages. Naseem et al. (2012); Zhang and Barzilay (2015) on the other hand looked at language typology features to improve the performance of model transfer. Many other work, as we will show more in detail later in the chapter, made use of cross-lingual word embeddings (e.g. Xiao and Guo, 2014; Guo et al., 2016) or multilingual word-clusters (Täckström et al., 2012). Another line of work using neural network based models have also tried to share the parameters between source and target language (Duong et al., 2015).

Beside work on cross-lingual dependency parsing, more recently Reddy et al. (2017) have attempted to convert UD annotation to logical form for universal semantic parsing and Pražák and Konopik (2017) have used UD annotations for cross-lingual SRL.

5.2.2 Tree-structred LSTM

Tree-structured LSTM (Tai et al., 2015, treeLSTM below) are *recursive* neural models that accept trees instead of linear sequences as input. In doing so, these architectures are able to encode syntactic (or similar hierarchical) proprieties underlying an input string while ignoring order-sensitive information.

We refer here to the original treeLSTM formulation which is *unidirectional*, i.e. an input tree is only traversed bottom-up. Unlike recurrent LSTM where the state of a time step is composed by weighing an input embedding and information from a previous time step, in a treeLSTM it is the input embeddings of the current node and the states of one or more of its children that are composed together. By applying this process recursively bottom-up we are able to propagate information until the root node is reached and an embedding encoding the entire input returned.

Tai et al. (2015) presents two variants of a treeLSTM architecture: a *child n-ary treeLSTM* and a *child-sum treeLSTM*. The former suits tasks where the input tree has a known constant branching factor; on the other hand, the latter suits tasks where the input tree has variable and often large branching factor. Since in this work we are going to use exclusively dependency structures, we will experiment with child-sum treeLSTM only; this is the model that we describe below.

A treeLSTM presents the same basic computation of a standard LSTM with an input and output gates i and o, a memory cell c and hidden state h. However, unlike its recurrent counterpart, the computation of the gates and memory cell for a unit has to account for possibly multiple children. This is the case of the forget gate f in particular, where it is desirable to weight separately for each children the amount of information to pass onto its parent state; this is achieved by having as many forget gates as the number of children.

The computation of the hidden state for a node j in the tree is computed in the same way as Eq. 5.1, where $C(j)$ is the set of children states of node j; x_j is the input embedding for node j; i_j, o_j, f_j are the input, output and forget gates; u_j an input tanh layer; c_j and h_j the memory cell and hidden state for the node j respectively.

As shown in Eq. 5.1 children are summed together when computing all gates except for the forget gate where the contribution of each child is computed separately and then summed together.

5.2.3 Graph Convolutional Networks

Graph Convolutational Networks (GCN) are a family of graph-based neural architectures. The intuition behind GCN is that the hidden representation for each node in the graph is a function that aggregates information from its immediate neighbors.

$$\begin{aligned}
\widetilde{h}_j &= \sum_{k \in C(n)} h_k \\
i_j &= \sigma(W^{(i)} x_j + U^{(i)} \widetilde{h}_j + b^{(i)}) \\
o_j &= \sigma(W^{(o)} x_j + U^{(o)} \widetilde{h}_j + b^{(o)}) \\
u_j &= tanh(W^{(u)} x_j + U^{(u)} \widetilde{h}_j + b^{(u)}) \\
f_{jk} &= \sigma(W^{(f)} x_j + U^{(f)} h_k + b^{(f)}) \\
c_j &= i_j \cdot u_j + \sum_{k \in C(n)} f_{jk} \cdot c_k \\
h_j &= o_j \cdot tanh(c_j)
\end{aligned} \tag{5.1}$$

To communicate information between nodes that are not immediate neighbors, this process is iterated a fixed number of times, where each iteration is implemented by a corresponding neural network layer. GCNs do not assume that their input directed, so they have no notion of bottom-up or top-down traversal; directionality is encoded explicitly into the neighborhood function. Furthermore, unlike treeLSTMs, a GCN lacks memory cells which might be a disadvantage when considering the problem of vanishing gradients.

Despite GCN having been initially formalized by Kipf and Welling (2016), we refer here to the model of Marcheggiani and Titov (2017) who applied it for the first time to an NLP task, semantic role labeling.

Similarly to a treeLSTM, a GCN takes as input an embedding for each node in the graph. These inputs is then passed onto a non-linearity or through a bi-LSTM before being fed to the GCN.

The computation for the hidden state of a given node v takes into account: the hidden state of a neighbor node n; the directionality of the edge between v and n and the dependency label with its directionality specified. For each directionality a different weight matrix W$_{dir(u,v)}$ is used; information regarding the dependency label is instead encoded in the bias vector b$^{l(u,v)}$. This yields the following equation:

$$\begin{aligned}
\mathbf{h_v^{(K+1)}} = ReLU(\\
\sum_{u \in \mathcal{N}(v)} g_{v,u}^{(K)} (\mathbf{W^{(K)}}_{dir(u,v)} \mathbf{h_v} + \mathbf{b^{l(u,v)}}))
\end{aligned} \tag{5.2}$$

where g$_{(v,u)}$ is an edge-wise scalar gate to help weighing the importance of an edge-node pair amongst several neighbors.

5.2.4 Cross-lingual word embeddings

In recent years, methods to create a cross-lingual embedding space have emerged and have been widely used in a variety of tasks requiring reasoning about word semantics across languages, including parsing (Ammar et al., 2016) and language understanding (Mrkšić et al., 2017). In this section, we refer to the survey of Ruder et al. (2017) that offer a high-level comprehensive view of the work in the field.

A first axis of variation amongst work on building cross-lingual embeddings lies in the type of bilingual data used. We can therefore distinguish between: *word alignment* methods, that leverage automatically extracted word alignment information or human-curated bilingual dictionaries to bridge in between languages; *sentence alignment* methods 'á la MT', where mapping between source and target language is extracted from parallel sentences; and *document alignment*, where this mapping is extracted from non-strictly parallel documents, such as Wikipedia pages in different languages. More in depth, each of these sub-categorizations can be characterized as follows.

Word-based alignment. Work leveraging a bilingual mapping between words differ both in the models used to bring together source and target embedding spaces, as well as in the type of alignment used.

A line of work have tried to learn a matrix transformation between the two spaces under the assumption that the geometry of the embedding space is similar across languages. Mikolov et al. (2013a) learns this weight matrix via SGD where the goal is to minimize the mean square error (MSE) over the euclidean distance between translation pairs in a given bilingual dictionary. A similar idea is also used in Dinu et al. (2014) and Faruqui et al. (2015), where Max-Margin Hinge-Loss and canonical component analysis(CCA) are used in place of MSE respectively.

Different work have also considered different alignment information. Whereas the aforementioned work used bilingual dictionaries, other work have experimented with 'pseudo-bilingual corpora', where either source and target words or their contexts are mixed together or where alternative representations are used. Xiao and Guo (2014) created a joint vocabulary out of Wiktionary where the same vector representation is assigned to each translation pair and the embedding space is trained by feeding both source and target context words. Gouws and Søgaard (2015) have experimented with a mixed corpus where source words are substituted with their target translation (or with a word with the same PoS) according to a given probability. Bergsma and Van Durme

(2011), on the other hand, have used images as language-agnostic data to bridge between source and target languages.

Sentence-based alignment. Similar to work leveraging word-level information, another line of research have looked into bilingual corpora aligned at sentence level to compute the embedding space. Some of these work operate under the assumption that the source and target sentence can be represented as the sum of their component words and a bilingual embedding space can emerge by minimizing the distance between these sums (Hermann and Blunsom, 2013, 2014). Others try to reconstruct the target sentence from the source, and viceversa by means of autoencoders (Lauly et al., 2014) or leverage skip-gram with negative sampling (Mikolov et al., 2013b) by assuming that the alignments between source and target sentence are uniform (Gouws et al., 2015) or monotonic (Luong et al., 2015). Finally, similar to the works discussed above leveraging images as language-independent signal, Calixto et al. (2017) and Gella et al. (2017) also used images and tried to bring captions and their translations closer together.

Document-based alignment. A final line of work uses document-aligned data to compute a bilingual distributed representation. Documents are advantageous in that they are cheaper to obtain; on the other hand, the sentences are not parallel and explicit word-alignment information is therefore missing. Nonetheless given the availability of large multilingual document-aligned resources (e.g. Wikipedia), a few works have attempted to leverage this information. Vulić and Moens (2016) for instance tries to mix the contents of source and target document by iteratively picking words from either to create a pseudo-bilingual corpus, which is then used to extract a word embedding space. On the other hand Vulić and Moens (2013) and Søgaard et al. (2015) base their methods on the assumption that words around a given topic or evoking the same concepts in both languages should be clustered around the same space in both languages; in particular, Søgaard et al. (2015) uses Wikipedia topics to test this assumption where a word is associated to the concept it describes.

It is worth noticing however that the ideas underlying bilingual embedding spaces described above can be already found in earlier 'pre-embedding' works. For instance, cross-lingual clusters (e.g. Täckström et al. (2012)) already described a method where a target word is assigned a target cluster (and therefore a position in the target embedding space) conditioned on word alignment information (or to be more precise, on the source cluster the aligned word belongs to).

5.3 The model

The detection task is formalized the same way we described in §3.2, that we summarize here again for convenience.

Given a sentence $w = w_1...w_{|w|}$, we encode the cue as a binary vector $c \in \{1,0\}^{|w|}$, where $c_i = 1$ if a token is part of the cue and 0 otherwise. Our goal is to predict the negation scope $s \in \{1,0\}^{|w|}$, where $s_i = 1$ if a token is part of the scope and 0 otherwise. All of our models are neural probabilistic models of $p(s_i|w,c)$, where w and c are encoded using different architectures, depending on whether we use a dependency graph as input. When we do not, the model is simply a BiLSTM over w and c.

We now turn to the encoding of dependency structures, considering the example in Figure 5.2, which is annotated as follows:

(90) <u>Sarah</u> went home and <u>was</u> **not** <u>seen since then</u>

The input can be (as in this case) a directed acyclic graph (DAG), so our model must account for this. We can traverse the graph bottom-up, from leaves to root, or top-down, from root to leaves. A top-down pass seems insufficient, since negation since cues are usually terminals as in the example. On the other hand, a bottom-up pass would capture the dependency chain ‘not’ $\rightarrow$ ‘seen’ but would miss the sibling tokens of ‘not’ that are in scope. Hence we need a bi-directional model that can encode the DAG bottom-up and top-down. But this is still insufficient unless the passes communicate: that is, if the bottom-up pass first collects information about the children of ‘seen’, then the top-down pass can pick up that information and pass it downward, hence communicating information about ‘not’ to its sibling nodes in scope. In both passes, the model should also be aware the same node can be different relations with multiple parents; this is the case of ‘Sarah’ in the example, which is the *subj* of ‘went’ but the *subj:pass* of ‘seen’.

Whereas Marcheggiani and Titov (2017)’s GCN meet these requirements, the unidirectional treeLSTM of Tai et al. (2015) falls short in that 1) it doesn’t accept DAGs as input and 2) it does not allow for bi-directional encoding. For this reason, while experimenting with the original GCN implementation, we enhance the original treeLSTM formulation to address these two shortcomings. The resulting bi-directional child-sum DependencyLSTM (BiDLSTM below, modeled after Tai et al. (2015)) is described in the following section.

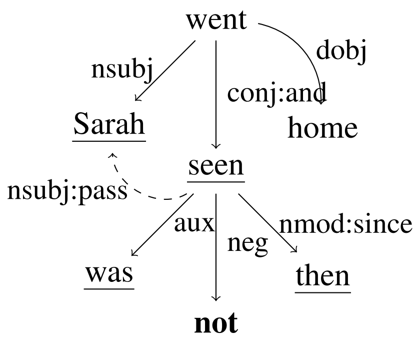


Figure 5.2: Dependency parse (using UD v1) for the sentence ‘Sarah went home and was not seen since then’, where the dashed line represent an enhancement to the parse available in the UD Enhanced++ representation.

5.3.1 Bi-directional Dependency graph LSTM (BiDLSTM)

Our model can accept as input both a tree or a DAG; we will refer to either as a dependency *structure*. A dependency structure g is a tuple (V,E), where V is the set of word-nodes v and E the set of dependency edges e. Each edge $e \in E$ is assigned a dependency label l. We define as $P(v)$ the set of parents of node v and $C(v)$ the set of its children; if the dependency structure is a tree $\mid P(v) \mid = 1 \; \forall v \in V$. We also define T as the set of terminal nodes and N as the set of internal nodes, where $T \subset V$ and $N \subset V$. r is the root node.

We represent both $v \in V$ and $e \in E$ as d-dimensional embedding vector, $\mathbf{w}$ and $\mathbf{l}$, where $\mathbf{w} \in \mathbb{R}^{\mathrm{d_w}}$ and $\mathbf{l} \in \mathbb{R}^{\mathrm{d_l}}$. We also encode information 1) about whether a word w is a cue or not, encoded in a cue-embedding vector $\mathbf{c} \in \mathbb{R}^{\mathrm{d_c}}$ and 2) about the universal PoS tag of w, represented as a PoS embedding vector $\mathbf{p} \in \mathbb{R}^{\mathrm{d_p}}$.

This information is encoded in an input vector $\mathbf{x_v}$ as follows:

$$\mathbf{x_v} = \mathbf{W}[\mathbf{x_t};\mathbf{c_t};\mathbf{p_t};\mathbf{l}_{(p_i^t,t)}] + \mathbf{b} \tag{5.3}$$

where all features are concatenated together and passed through a first transformation to match the dimensionality of the hidden layer. This allows multiple layers to be stacked together, where $\mathbf{x_t}$ can be replaced with the output hidden state of the previous layer.

To address the lack of support for DAGs where a node might have multiple parents, we create as many states as the number of its parents; given a node v, we define this set as $S(v)$. This layout can also reflect that the same node may be connected to different parents with different labels; for instance, if the dashed line in Fig. 5.2 had the label 'nsubj:pass' we would have nedeed two separate representations for the word 'I', one where l='nsubj' and one where l='nsubj:pass'.

During a first bottom-up we traverse the nodes in reverse topological order; the computation is the same as Eq. 5.1, except that instead on the input $\mathbf{x}$ we pass $\mathbf{h}_\mathbf{v}^{\uparrow}$ and $\mathbf{c}_\mathbf{v}^{\uparrow}$. This pass returns the hidden states and the memory cell for the root node, $\mathbf{h}_\mathbf{r}^{\uparrow}$ and $\mathbf{c}_\mathbf{r}^{\uparrow}$, which for convenience we represent together as the state $\mathbf{s}_\mathbf{v}^{\uparrow}$

To address the lack bi-directionality in the original child-sum TreeLSTM of Tai et al. (2015), we add a second top-down pass where we feed the states computed during the bottom-up pass; in this our model is very similar to the one of Chen et al. (2017) and to the inside-outside NN architecture of Le and Zuidema (2014).

The top-down pass is similar to the bottom-up one but traverses the vertices in a topological order. Again, for a given node, we compute as many states as the number of its parents. To create a dependency between passes, we also made the states computed during the bottom-up pass, $\mathbf{s}_\mathbf{v}^{\uparrow}$, available in the form of additional weighted feature during the top-down pass. We start by computing the representation of the root node r as follows:

$$\mathbf{s}_\mathbf{r}^{\downarrow} = LSTM(\mathbf{x}_\mathbf{r}, \mathbf{s}_\mathbf{r}^{\uparrow}) \tag{5.4}$$

When computing the state of a node top-down, we sum up the parent state(s) the same we did for the children state(s) in the bottom-up pass. However, since we create as many states as the number of parents, the parent node itself and the current node's bottom-up representation might have be composed of different states; if this is the case, we sum these states. The hidden representation of the remaining nodes v, $s_v^{\downarrow}$, is computed as follows:

$$\begin{aligned}
\mathbf{s}_{\mathbf{p(v)}}^{\downarrow} &= \sum_{j \in |S(p(v))|} \mathbf{s}_j \\
\mathbf{s}_\mathbf{v}^{\uparrow} &= \sum_{j \in |S(v)|} \mathbf{s}_j^{\uparrow} \\
\mathbf{s}_\mathbf{v}^{\downarrow} &= LSTM(x_v, s_v^{\uparrow}, s_{p(v)}^{\downarrow})
\end{aligned} \tag{5.5}$$

After both passes are computed, we pass the hidden states obtained at the end of the top-down pass to the softmax layer to compute the probability of a given node to be inside or outside the scope of negation.[1] Since for one node there are as many hidden

[1] We also experimented with concatenating the two passes together but saw no difference in performance

representation as the number of parents, we take the sum of these. The computation is as follows:

$$h_v = \sum_{j \in S(v)} h_j^{\downarrow}$$
$$\hat{p}(y|h_v) = softmax(\mathbf{W h_v} + \mathbf{b}) \tag{5.6}$$

Figure 5.3 illustrates the computation of the Dependency LSTM and to the one performed by GCN.

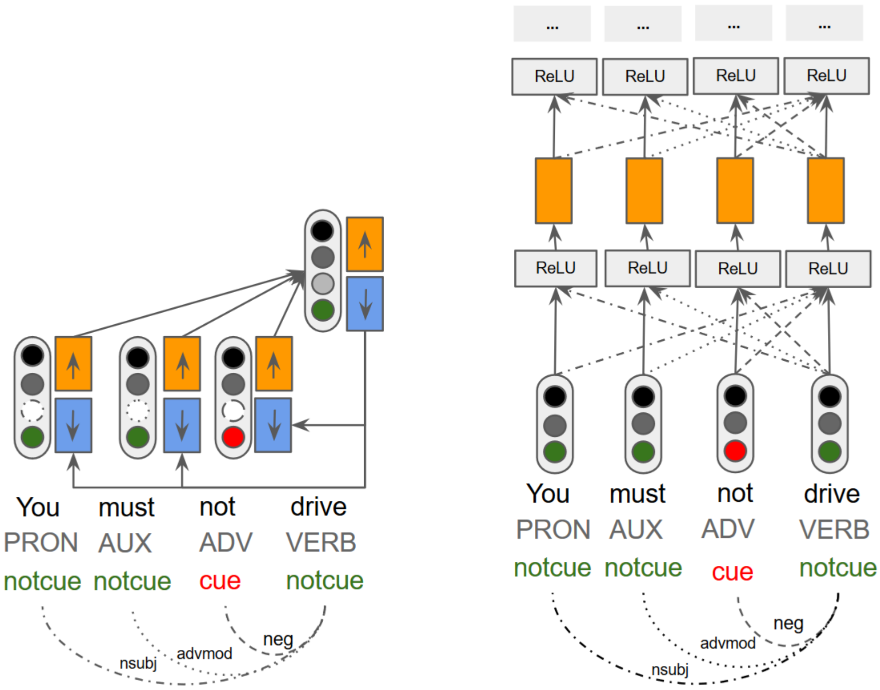


Figure 5.3: Computation flow for the Bi-DLSTM architecture(left) and the GCN (right) for the sentence 'I must not drive'. In the Bi-DLSTM each word is represented by the concatenation of word, universal PoS tags, dependency label and cue features. The bottom-up pass builds from the leaves ('you', 'must' and 'not') to the root ('drive') and the top-down in the opposite direction. The states built during both passes are exemplified by the ↑ and the ↓ respectively. In the GCN, hidden representations are built by aggregating neighboring nodes in the dependency trees, as represented by the dashed lines. The node itself is also taken into consideration as shown by the straight lines. Information propagates by stacking up different layers.

5.4 Data and experiment settings

To train and test our model, we split NEGPAR into train, development and test sets following the same split as CONANDOYLENEG. Statistics are reported in Table 5.1.

	English	Chinese
train	981	1206
dev	174	230
test	263	341

Table 5.1: Number of negation instances in the train, dev and test set in the English and Chinese sides of NEGPAR

We assess the validity of our method first on both English and Chinese separately and then across language by training in English and testing in Chinese. For both English and Chinese, we obtain PoS tags and UD v1 parses using the Stanford Parser (Chen and Manning, 2014); PoS tags are then converted into universal PoS tags.[2]. The word-segmentation the Chinese side of NEGPAR is based on also leverages Stanford toolkits (Chang et al., 2008).

Despite the models are already agnostic to node ordering, we also remove any punctuation (and related edge) from the dependency structure. Moreover when testing cross-lingually we remove language-specific dependency tags (e.g.conj:and→ conj).

We use cross-lingual word embeddings pre-trained on Wikipedia data[3] where a linear transformation has mapped Chinese and English embeddings into a common embedding space; in a English-to-Chinese lexical similarity task the method has a p@1 of 0.40 and a p@5 of 0.68(Smith et al., 2017).

We explore the following experimental settings:

- **UD v1 vs. UD++ vs. v2**: one difference in between UD v1 and v2 is the replacement of the *neg* label with an *advmod* relation, whereas information on the polarity of a token is included in its morphological features. We assess whether this difference, among others, has an impact on our model. For both English and Chinese, UD v1 parses are converted into v2 using the official UD conversion tools[4]. Since we are dealing with long-range dependencies we also experiment

[2] Mapping available at `https://github.com/slavpetrov/universal-pos-tags`

[3] Available at `https://github.com/Babylonpartners/fastText_multilingual`

[4] `http://universaldependencies.org/tools.html`

with the Enhanced++ version of Stanford Dependencies (Schuster and Manning, 2016).

- **Pre-trained word embeddings**: given the small amount of training data, we assess whether using pre-trained word-embeddings improves performance in the monolingual English and Chinese settings. To this end, we use the pre-trained word embeddings mentioned above but where no linear transformation is applied.

- **Feature ablation**: to assess the contribution of word, dependency label and universal PoS tag we carry out a feature ablation study. We consider here dependency labels as indispensable for the task and experiment with removing the word-embedding feature first (*-w*) and then the PoS-embedding (*-p*).

Hyperparameter tuning was performed separately for both the monolingual and the cross-lingual task, for each language and for each UD version. The hyperparameters used are shown in Table 5.2 and 5.3 for the monolingual and the cross-lingual settings respectively.

	BiDLSTM					GCN				
	en			zh		en			zh	
	UD1	UD2	UD++	UD1	UD2	UD1	UD2	UD++	UD1	UD2
d_h	300	300	200	200	300	300	300	300	400	300
d_w	50	100	50	50	50	50	50	200	50	50
d_c	20	30	30	20	1	20	30	20	10	10
d_p	100	100	200	50	100	100	50	50	200	200
d_l	100	50	100	100	200	200	50	50	100	100
dropout	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1

Table 5.2: Hyperparameters used in the monolingual experiments.

We optimize both models using Adam(Kingma and Ba, 2014), with an initial learning rate of 0.01. In the BiDLSTM dropout is performed on the output layer, whereas in the GCN we follow Marcheggiani and Titov (2017) in performing dropout on the neighbors $N(v)$.

We evaluate the model in the same way described in Ch 3 by reporting token-based precision, recall and F_1, as well as the % of scope spans correctly and fully detected (PCS). All results are reported as an average of 5 runs.

	BiDLSTM		GCN	
	UD1	UD2	UD1	UD2
d_h	400	300	300	300
d_c	30	30	10	10
d_p	50	50	200	50
d_l	50	50	50	50
dropout	0.2	0.2	0.2	0.1

Table 5.3: Hyperparameters used in the cross-lingual experiments. A fixed word embedding dimension of 300 was used to match the dimensionality of the pre-trained cross-lingual word embeddings.

5.5 Results

5.5.1 Monolingual

Which model performs better in a monolingual setting? Results on the English and Chinese *dev* sets are shown in Table 5.4. The BiDLSTM model outperforms the GCN in all settings considered and in both languages; differences for both token-level F_1 and PCS are statistically significant. In English, we found a significant difference in terms of F_1 scores between the BiDLSTM model trained on UD++ parses and the other two models; however, we did not find any significant different between UD versions for the GCN models. In Chinese, we found no statistical difference between the F_1 scores of the BiDLSTM trained on different UD versions ($p = 0.57$). As for the GCN, the difference between UDv1 and UDv2 is statistically significant at $p < 0.05$. Finally, we also did not find any significant difference in using pre-trained word embeddings.

Are some features more relevant than others? Results for the feature ablation experiments are reported in Table 5.5 for English and Chinese respectively.

For the BiDLSTM, only removing word embeddings leads to a statistically significant drop in performance ($p < 0.01$) in terms of F_1 measure. For the GCN, on the other hand, we did not observe any significant difference when removing either word-embedding or PoS-embedding feature. However, the difference in performance between including and excluding word embeddings as features is not substantial, suggesting that a structured network is guided in its predictions mostly by information available on the dependency parse.

How does a recursive model compare against a recurrent one? We compare the

	BiDLSTM				GCN			
	P	R	F_1	PCS	P	R	F_1	PCS
UD1	83.10	85.66	84.23(±0.75)	47.88 (±1.07)	83.84	78.14	80.86(±1.08)	38.81(±1.51)
UD2	82.55	**87.45**	84.92(±0.92)	**49.56**(±1.27)	81.94	78.60	80.06(±1.42)	37.68(±1.69)
UD++	**85.28**	86	**85.59**(±0.61)	48.39(±0.41)	80.91	80.81	80.84(±0.75)	36.74(±1.35)

(a) English

	BiDLSTM				GCN			
	P	R	F_1	PCS	P	R	F_1	PCS
UD1	**79.99**	65.76	**72.13**(±1.10)	25.80(±0.85)	70.37	65.28	67.68(±1.41)	19.57(±1.01)
UD2	76.15	**67.71**	71.60(±1.25)	**26.72**(±0.47)	74.75	64.57	69.24(±0.60)	21.06(±1.76)

(b) Chinese

Table 5.4: Results for the monolingual setting on English and Chinese respectively when different UD versions are compared. Best results are reported in bold for each language.

performance of the BiDLSTM against a BiLSTM, which were shown to yield to state-of-the-art performance for this task. Results are shown in Table 5.6. In both languages we found the BiDLSTM to perform worse than the BiLSTM we previously developed both when punctuation is included or when it is removed.

5.5.2 Cross-lingual

What's the performance of our models across languages? Results on the Chinese *dev* set are reported in Table 5.7. The BiDLSTM outperforms once again the GCN model; results for both UD versions are statistically significant ($p < 0.01$). However for the BiDLSTM model we did not find a statistically significant difference between the two UD models ($p < 0.20$).

Results on the *test* set are reported in Table 5.8, where the performance of the BiDLSTM and the GCN is compared to the BiLSTM model, with and without punctuation. As expected, the structured models outperforms the recurrent models by a large margin both in terms of token-based F_1 measure and PCS. Whereas recurrent models are state-of-the-art in a monolingual setting, the lack of structural information make them unsuitable for cross-lingual negation scope detection.

Do the features contribute the same way as in the monolingual task? Results are shown in Table 5.9. When removed, words embeddings and universal PoS tags do not lead to a statistically significant drop in performance ($p > 0.05$), highlighting again that also in the cross-lingual task the system relies mostly on information coming from

		P	R	F_1	PCS
BiDLSTM (UD++)	*all*	**85.28**	86	**85.59**(±0.61)	48.39(±0.41)
	-w	81.02	**86.52**	83.64(±0.55)	**51.39**(±0.53)
	-p	84.26	84.64	84.40(±1.20)	48(±0.84)
GCN (UD1)	*all*	**83.84**	78.14	80.86(±1.08)	38.81(±1.51)
	-w	82.89	79.27	**81.01**(±0.29)	**42.93**(±0.49)
	-p	80.96	**80.54**	80.65(±1.28)	41.77(±0.76)

(a) English

		P	R	F_1	PCS
BiDSLTM (UD2)	*all*	76.15	**67.71**	**71.60**(±1.25)	**26.72**(±1.47)
	-w	75.01	64.88	69.56(±0.51)	21.31(±0.50)
	-p	**77.95**	65.80	71(±1.46)	22.36(±1.42)
GCN (UD2)	*all*	74.75	64.57	69.24(±0.60)	21.06(±1.76)
	-w	75.14	63.92	69.04(±0.28)	20.81(±0.32)
	-p	**77.95**	**65.80**	**71.00**(±2.46)	**22.36**(±0.89)

(b) Chinese

Table 5.5: Results from the feature ablation studies for the best systems in both English and Chinese. *-w*: word embedding feature removed; *-p*: universal PoS embedding feature removed

the dependency trees and less on lexical information. Moreover, the PCS for both systems seem to be higher when removing these two features. These results apply to both structured models.

For completeness, we also assess the role of the word embedding feature for the BiLSTM models, shown at the top of Figure 5.4. These results show two different aspects: the first, and more obvious, is that recurrent models overfit to to the training data, with a great loss in performance when moving from *dev* to *test* set. The second and more surprising finding is that the performance is on par with our best structured model. We will come back to this in the next section.

5.6 Discussion

Given these results, we felt three aspects needed to be explored more in depth.

Parse structure vs. cross-lingual word embeddings Results in Table 5.9 has shown

		P	R	F_1	PCS
en	BiLSTM	**88.39**	**87.16**	**87.71**(±0.65)	**53.54**(±0.71)
	BiLSTM(no punct.)	85.21	83.56	84.32(±0.38)	51.35(±1.17)
	BiDLSTM	85.28	86	85.59(±0.61)	48.39(±0.41)
zh	BiLSTM	**83.73**	**73.58**	**78.18**(±0.28)	**33.04**(±0.98)
	BiLSTM(no punct.)	74.55	73.18	73.82(±0.94)	26.76(±0.42)
	BiDLSTM	76.15	67.71	71.60(±1.25)	26.72(±1.47)

Table 5.6: Comparison between our best recursive model and the BiLSTM model trained on data with (*BiLSTM*) and without punctuation (*BiLSTM(no punct.)*

		P	R	F_1	PCS
UD1	BiDLSTM	**66.98**	67.92	**67.11**(±1.85)	**10.33**(±1.39)
	GCN	53.43	**70.38**	60.57(±1.04)	8.75(±1.41)
UD2	BiDSLTM	61.22	70.09	65.15(±2.45)	8.88(±1.11)
	GCN	61.10	58.69	59.75(±2.04)	4.26(±1.08)

Table 5.7: Results for the cross-lingual detection task on the Chinese *dev* set

how structural information can already by itself encode information regarding negation scope, with little help from cross-lingual word embeddings. However why are cross-lingual word embeddings not helpful in detecting negation? This question is hard to answer but we hypothesize this happens for two reasons.

A first reason may be that lexical semantic similarity is not a representation that the models find useful when detecting negation scope. Although it is safe to assume that what drives prediction in both the BiDLSTM and the GCN model is the UD parse, it is not clear what does in the BiLSTM models, where performance on the *dev* set is still competitive even in the absence of UD trees or word embedding information as input.

We hypothesize once again this has to do with negation scope boundaries, as we saw in Chapter 3 and we investigate once again the ability of the model to predict easy vs. hard cases. As a reminder, this is exemplified in below, where easy cases are those where negation scopes can be predicted by including all the tokens to the left and right of the cue up to the first punctuation or sentence boundaries whereas the hard ones are those where we cannot.

(91) *Easy*: “You are **not** ready” , she told me
Hard: She did **not** said anything because it was too late

	P	R	F$_1$	PCS
BiLSTM	56.49	29.97	38.69	3.97
BiLSTM(no punct.)	51.54	26.76	31.55	3.65
BiDLSTM (UD1)	**68.45**	65.35	**66.56**	**14.37**
GCN (UD1)	55.50	**72.34**	62.61	10.22

Table 5.8: Results for the cross-lingual detection task on the Chinese *test* set

		P	R	F_1	PCS
BiDLSTM (UD1)	*-all*	66.98	67.92	**67.11**(±1.85)	10.33(±1.39)
	-w	60.80	**71.41**	65.47(±1.02)	**14.80**(±0.22)
	-p	**69.64**	62.48	65.62(±1.07)	14.60(±1.52)
GCN (UD1)	*-all*	53.43	**70.38**	60.57(±1.04)	8.75(±1.41)
	-w	**62.31**	63.35	**62.81**(±0.88)	**12.32**(±0.28)
	-p	59.26	62.87	60.85(±1.61)	10.12(±0.63)

Table 5.9: Results of the feature ablation experiment using the best performing cross-lingual models

In doing so, we verify whether this trend holds for the BiDLSTM and the GCN model as well.

Figure 5.5 plots the % of easy cases vs. the % of hard cases correctly predicted by the best systems for the four models considered.

In the case of the BiLSTM where punctuation is kept the gap is visible in both monolingual settings confirming the results of Chapter 3. To further prove this is the case, we looked into the hard cases that were not predicted correctly and we analyzed the words immediately preceding and following the scope span predicted by the the BiLSTM. We found that 29% of these words are full stops and 23% are commas, showing that in more than half of the cases scope prediction is driven the presence of a punctuation before and after a given cue. The system learns in fact to detect as scope everything from the cue up to the first instance punctuation or sentence boundary, as shown in the example below (system predictions are reported in curly brackets{}).

(92) 自己 颇 不 {寻常 的 外表} 。
own rather not common DE appearance .
'...[his] own rather unconventional appearance.'

However, punctuation alone cannot explain the results for the BiLSTM model where

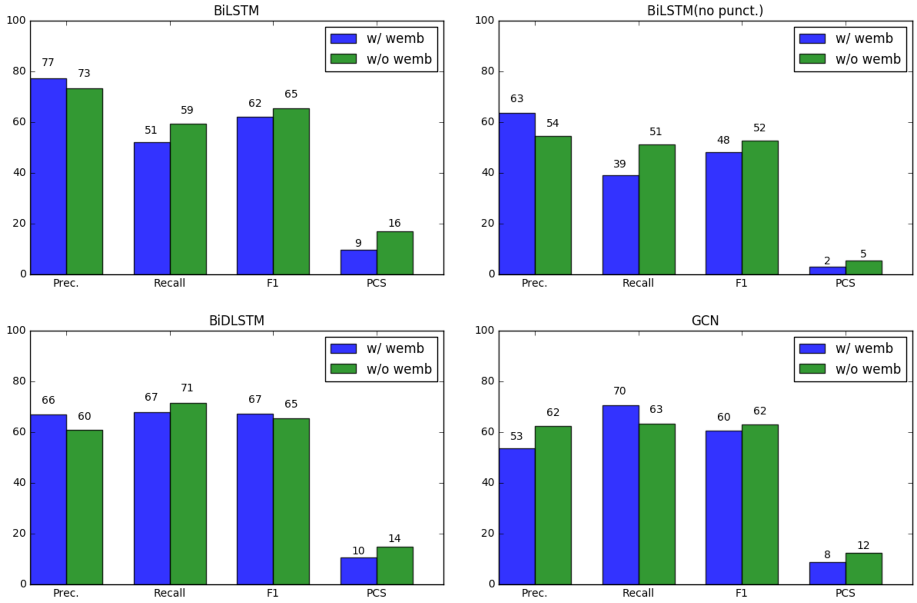


Figure 5.4: Performance of the models training with and without cross-lingual embeddings.

it was removed. We looked both at the hard cases that were correctly and incorrectly predicted and found that this models relies on the fact that certain some syntactic environments exactly matching the scope are marked consistently. For instance, out of 44 hard cases correctly predicted for English we found 31 cases where the model used the conjunction ’and’ and ’but’, as well as the relativizer ’that’ to mark the boundaries for predictions.

For the BiDLSTM and the GCN, we notice that the difference between easy and hard cases correctly predicated is smaller when compared to the BiLSTM. However, English seem to be the exception; we will come back to this point in the next section.

Overall these results show that **negation scope detection is indeed a task involving mostly structural information and less lexical information. However, what kind of “structural information” is used depends from the model: in structured models this information is explicit given the input parse and that is what model learns to use. In recurrent models, where this information is not explicitly given, structure is found in the form of boundaries around sentence spans.**

We also hypothesized that word embeddings might benefit from fine-tuning during training; however, even fine-tuning them does not lead to any improvement, with a drop in performance of 2.5F_1 points on avg. for all systems.

The quality of the cross-lingual word embeddings themselves might play a role but

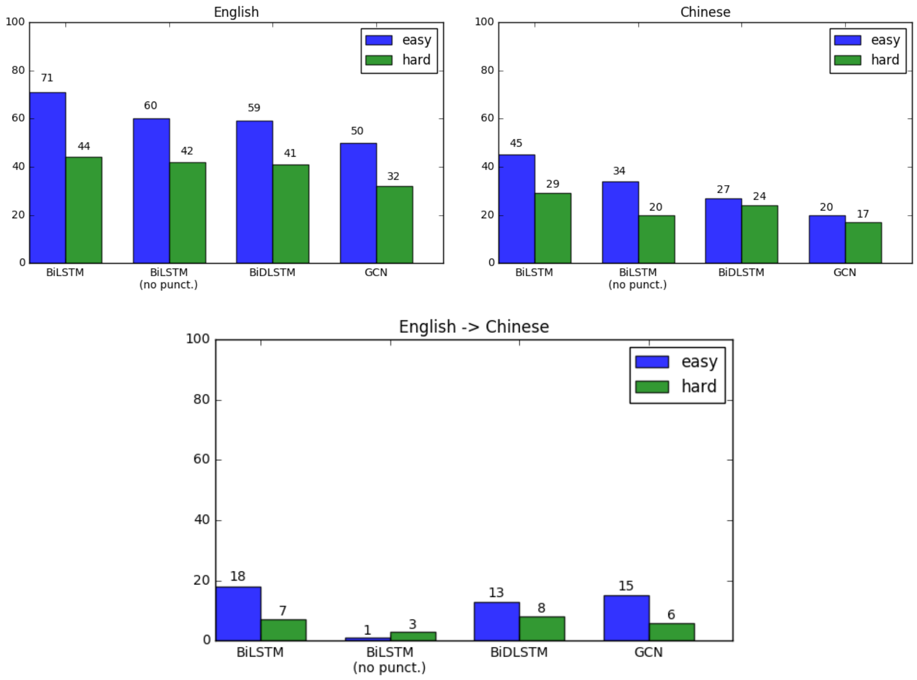


Figure 5.5: Performance of the models on the easy and hard cases for the three models here considered (BiLSTM, BIDLSTM, GCN) across the monolingual and the cross-lingual task

this point is hard to verify since it can only be evaluated extrinsically. For instance, Smith et al. (2017) has used this same word embeddings for the task of lexical translation: where given a word in English, e.g. ”dog”, the task is to find a word in the target language whose cosine similarity is the highest. When predicting a Chinese word given the English, the method achieves a precision@1 is 0.40 and precision@5 is 0.68, meaning that more than half of the time the correct translation is in an n-best list of 5 most similar words proposed by the model. However, the quality of translation task might not have any bearing on the task of detecting detecting. Finally, it is worth mentioning that the pre-trained embedding only cover around 65% of the words in the training (as opposed to 85% for English); for the reminder of the words the embeddings are randomly initialized (and not updated).

Are the networks not initialized properly? Marcheggiani and Titov (2017) have shown that GCNs benefit greatly by the presence of a first BiLSTM layer. We hypothesize that linear information might help initialize the network better and we assess the impact of adding an LSTM layer to the best systems in each settings.

Results in Table 5.10 show that initializing our models does yield better results in

	en				zh				en→zh			
	P	R	F_1	PCS	P	R	F_1	PCS	P	R	F_1	PCS
BiDLSTM	85.28	86	85.59	48.39	76.15	67.71	71.60	**26.72**	66.98	**67.92**	**67.11**	**10.33**
BiDLSTM (+LSTM)	**85.33**	**86.85**	**86.04**	**52.09**	**77.65**	**73.09**	**75.05**	25.79	**68.83**	57.14	61.65	5.78
GCN	83.84	78.14	80.86	37.68	74.75	64.57	69.24	21.06	53.43	**70.38**	**60.57**	8.43
GCN (+LSTM)	**86.73**	**80.57**	**83.52**	**45.40**	**75.83**	**74.89**	**75.20**	**31.37**	**60.07**	60.43	60.21	**9.46**

Table 5.10: Comparison between initializations for the both the BiDLSTM and the GCN, compared with a BiLSTM without punctuation.

the monolingual setting alone, where the improvement is statistically significant for both the BiDLSTM and the GCN models. However, this is not the case in the cross-lingual setting, where we shown the cross-lingual embeddings to hinder the learning of the BiLSTM layer.

How much does the number of layers affect the performance? Another finding of Marcheggiani and Titov (2017) is that the number of layers affect the performance of GCNs. This is not a surprising results since graph networks need multiple passes for the information to be spread properly throughout the dependency structure. We therefore ran some experiment to assess the impact of the number of layers on both the GCN and the BiDLSTM. Results shown in Fig. 5.6 show that whereas the number of layers has a significant impact of the GCNs, but it slightly affect the performance of the BiDLSTM only in the cross-lingual and not in the monolingual task.

Is the comparison between BiDLSTM and GCN unfair? As shown in § 3.2, the GCN model encodes information about the dependency labels in the bias term. However, given the importance of this feature, we hypothesize the BiDLSTM may have an advantage in that the dependency label embedding is weighted alongside other input features. To address this imbalance, we shift dependency label information from the bias to a dedicated weighted term, resulting in Eq 5.7

$$\mathbf{h_v^{(K+1)}} = ReLU\Big(\sum_{u\in\mathcal{N}(v)} g_{v,u}^{(K)}\big(\mathbf{W^{(K)}}_{dir(u,v)}\mathbf{h_v} + \mathbf{W_l^{(K)}}\mathbf{l}_{(u,v)} + \mathbf{b}\big)\Big) \tag{5.7}$$

The comparison between the vanilla GCN system and this modification (here referred as GCN+) in Table 5.11 shows that a GCN benefits from weighing the dependency label as a separate feature, with the difference being statistically significant ($p < 0.01$). However, we still found a statistically significant difference between the GCN+ model and the BiDLSTM ($p < 0.05$).

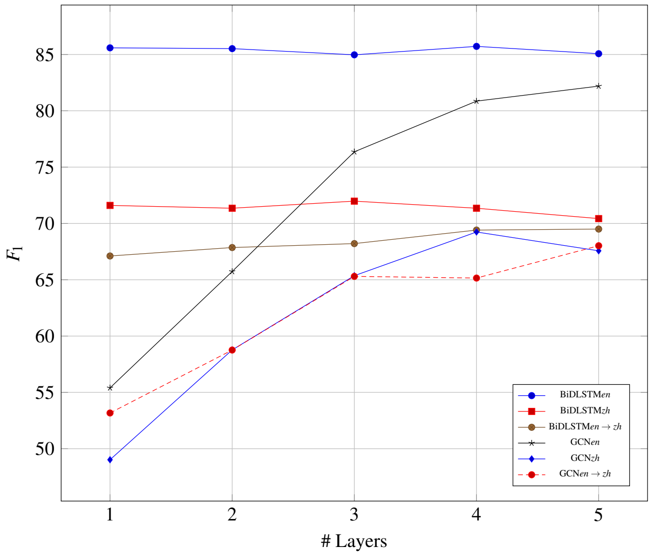


Figure 5.6: Number of layers plotted against the token-level F_1 for both the BiDLSTM and GCN model using UD v2.

	P	R	F_1	PCS
GCN	53.43	**70.38**	60.57	8.75
GCN+	**62.66**	65.67	**64.01**	**9.83**

Table 5.11: Comparison between the original GCN formulation of Marcheggiani and Titov (2017) and the one described in Eq 5.7 where the dependency label is a term weighed separately

5.7 Error Analysis

What is our model learning? We first analyze the performance of our best structured models by looking at the dependencies around negation scope. We take into consideration the parent edge of the least common ancestor for all the nodes in the scope; if the scope is discontinuous, we take into consideration the labels on top of all spans. For instance, in Fig. 5.2 the scope is discontinuous (if one doesn't consider the enhancement of the dashed edge) and it yields two sub-structures, one over the span 'was not seen since then', and one including the word 'Sarah'; we consider here the labels 'nsubj' and 'cc:and' as syntactic.environments.

In Table 5.12 we compare the performance of our best system, the BiDLSTM(UD1), across the two monolingual and the cross-lingual settings. In English we found that high performance is related to how well defined the syntactic environment around negation

en			zh			en→zh		
label	F_1	PCS	label	F_1	PCS	label	F_1	PCS
root	86.78	50	root	69.83	21.33	root	65.13	16
ccomp	74.11	44.82	conj	71.12	21.73	conj	66.10	13.04
conj:and	76.03	42.30	ccomp	72.86	30.55	ccomp	66.35	16.66
nsubj	69.67	28.57	nsubj	64.98	2.9	nsubj	55.71	0
nmod:of	88.23	42.85	dep	70.53	26.92	dep	59.34	11.53
advmod	76.92	28.57	dobj	73.73	12	dobj	61.94	0
xcomp	82.64	0	nmod:prep	73.47	10	nmod	55.90	5.2
dep	70.31	14.28	compound:nn	68.96	11	advmod	39.62	0
nmod:in	66.66	25	acl	45.16	11.11	compound	57.14	0
amod	96	50	advmod	60.75	0	acl	57.83	11

Table 5.12: Analysis of the syntactic environment around the scope where the dependency label represents the parent of the least common ancestor of all the nodes in the scope. Results are reported for the best BiDLSTM system (UD1) in the two monolingual settings (*en* and *zh*) and in the cross-lingual setting ($zh \rightarrow en$. Labels are ordered from most to least frequent.

scope is. In English, this is the case when negation scope spans the entire sentence (‘root’) since it is enclosed within sentence boundaries as well as when it spans a clausal complement (‘ccomp’) or a coordinate clause(‘conj’ and ‘conj:and’), that are introduced by specific marker (respectively ‘that’ and ‘and’ in this case). This can also be explained in conjunction with the experiments in the previous section where we assessed the performance of the model in predicting easy vs. hard cases. For instance, ‘root’ makes up 30% of all scope environments and it is often considered easy to detect since it is when negation scope matches sentence boundaries.

However a drop in performance for Chinese cannot be explained similarly, since neither relative not coordinated clauses are introduced by specific markers. What separates ‘root’, ‘ccomp’ and ‘conj’ from other dependency substructures is the fact that the cue is in the subtree as the scope. This is the case of the Figure 5.2 for instance where the cue 不 is the child of verb 喝 at the root of the coordinated construction ‘conj’.

In a cross-lingual settings, we noticed that for most cases there is a substantial loss in performance in terms of PCS but not in terms of F_1, meaning that although the scope is not exactly captured the model is still able to detect tokens within it.

In Table 5.13 we compared the performance of the BiDLSTM model against the best

GCN model (UD++) in the cross-lingual setting. Although prediction is even across all most frequent environment, the GCN fails to predict the same amount of full scope even in well defined environments.

en→zh

BiDLSTM(UD1)			GCN(UD1)		
label	F_1	PCS	label	F_1	PCS
root	65.13	16	root	55.61	6.75
conj	66.10	13.04	conj	61.26	4.44
ccomp	66.35	16.66	ccomp	60	8.5
nsubj	55.71	0	nsubj	60.5	9
dep	59.34	11.53	dep	60.82	7.69
dobj	61.94	0	dobj	57.97	12
nmod	55.90	5.2	nmod	61.27	5
advmod	39.62	0	advmod	60.57	9
compound	57.14	0	compound	44	0
acl	57.83	11	acl	50.90	11

Table 5.13: Analysis of the syntactic environment around the scope where the dependency label represents the parent of the least common ancestor of all the nodes in the scope. Results are reported for the best BiDLSTM and CGN model (UD1 and UD++ respectively) in the cross-lingual setting (*zh* → *en*). Labels are ordered from most to least frequent.

We then conducted a manual error analysis of the errors for the BiDLSTM to investigate whether there any patterns worth highlighting.

We found that all instances of neg. raising were not correctly captured by the systems. This is exemplified in (93.b) and (93.b) for the verbs 想到('to think') and 认为('to believe')

(93) a. {我 倒} 没有 {想到 你 身上 还 有 神经}
I instead have not thing you on your body still have strength
I did not think you still had strength

b. {我} 并不 {认为 我 已 弄清 全部 情况}
I not believe I already clarify all facts
I really don't think I have clarified all the situation already

On the other hand we found in 8 out of 10 instances to universal classifier was correctly predicted with respect to the scope of negation, as shown in the example below (94).

(94) {礼靴 和 背心 的 钮扣 都} **没有** 扣好}
Boots and vest DE button all not buttoned up properly
Boots and vest were both not buttoned up properly

Finally we found 8 cases where the systems does not distinguish the homographs 没有‘have not’, where both characters are part of the cue and 没有‘does not exist’, where only the first character is the cue and the second is the existential verb ’there is’ which is part of the scope. In these cases, the systems always include 没有as part of the scope as shown in (95).

(95) 要是 {我 没有 弄错} 的话 我们 的 当事人 已经 来 了
if I have not make a mistake if we DE interested party already come ASP
If it wasn’t for my mistake our person of interest would have come already

5.8 Chapter Conclusions

Let us go back to our initial research question: when detecting negation scope in a language other than English, *can we build a system to detect negation scope in a language with no annotations?* Yes. Although not as effective as an oracle model trained on that language, we show that this is indeed possible by means of a common syntactic annotation scheme, Universal Dependencies, and a neural network classifiers that recurses through an input parse.

We show however that cross-lingual word embeddings do not help or hinder the performance of our classifier. We hypothesize this is due to either the task requiring mostly syntactical than lexical information as well as to the quality of the word embeddings themselves.

Finally, through an error analysis we show that classification performance is higher when the cue is in the substructure as its scope, whereas our models fails in capturing those scope where lexical information is required.

Chapter 6

Conclusion

This chapter is based in part on the following peer-reviewed paper:

Federico Fancellu, Siva Reddy, Adam Lopez and Bonnie L. Webber(2017), *Universal Dependencies to Logical Forms with Negation Scope*, Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR), pages 22–32.

Throughout this thesis we have explored problems related to negation scope in a multilingual perspective. Our work was mainly motivated by the fact that previous work on automatically processing negation is mostly limited to English, with no exploration of how languages vary in representing negation and how models should be developed accordingly.

Our main contributions lie in detecting negation scope from corpora annotated at a string level; to this end, we investigated models that can be generalized across languages both in presence and in absence of annotated data for a language other than English, Chinese.

In doing so, we also reasoned more about the data, what are the effects it has on the performance of these models and the challenges involved in annotating negation at a string level in English and Chinese.

6.1 Main findings

We summarize the main findings of this thesis by considering the two scenarios considered in the task of negation scope detection.

Annotated data is available in both Chinese and English. This was the content of

Chapter 3.

For the task of *monolingual* negation scope detection, BiLSTM are state-of-the-art models that can be generalized across language. We have first explored the possibility of building a model achieving similar or better performance than previously developed classifiers, with a set of features available for languages other than English. We found bi-directional Long-Short Term Memory (BiLSTM) networks to satisfy our desideratum, outperforming previous work when tested on the same genre of data they are trained on, in presence of only word and universal PoS embedding features.

The performance of a BiLSTM model suffers from genre effects and differs from the type of cue triggering the scope. Furthermore, in most corpora annotated for negation scope, punctuation boundaries are already a strong baseline for negation scope detection. We also found punctuation boundaries to guide prediction in the BiLSTM model. Finally, adding a transition-based component on top of a BiLSTM helps in predicting continuous scopes.

We tested the robustness of our system by conducting two additional experiments.

We first tested on data of a different genre – sentences from Simple Wikipedia that we annotated from negation cue and scope. Results show performance to be affected by genre effects when compared to a non-neural classifier using syntactic features. We also show that performance is affected by the syntactic environment negation scopes on, with morphological negation on adjectives and adverbs being harder to capture than lexical negation, both verbal and non-verbal.

We also tested our system on all corpora available annotated for negation scope. Results have shown that BiLSTM models are consistently state-of-the-art for the task of detecting negation scope, with the proviso that enough training data is provided. We also found that for those corpora where negation is annotated as a continuous span of text, making the output predictions dependent on each other help performance. However, at closer inspection, we also found that for most corpora high performance can be attributed to a single property: negation scope is annotated as a single span of text delimited by punctuation or sentence boundaries. For negation scope not of this form detection accuracy is low and under-sampling the easy training examples does not substantially improve accuracy. We also demonstrate that this is partly an artifact of annotation guidelines.

Annotated data is not available in Chinese. Chapter 4 and 5 explored model transfer to deal with this scenario.

A parallel corpus annotated for negation allows to explore for divergences in representing negation across language and can serve as data for a cross-lingual model. In Chapter 4 we describe NEGPAR, a parallel English-Chinese corpus annotated for negation. The creation of NEGPAR is motivated by the fact that existing corpora in English and Chinese are not annotated in the same way for negation scope, reason why they are not a good fit for testing model transfer.

To ease the annotation process, we experimented with word-alignment based annotation projection, where we project existing English annotations of a set of stories from Sherlock Holmes onto their Chinese translations that we then manually correct.

However, prior to projection, we felt that the English annotations did not sufficiently brought to light or failed to capture at all some phenomena related to negation scope. For this reason, we reread the original annotation guidelines for English and reannotated some of the cases we thought should have been handled differently. In doing so, we also designed some additional guidelines to handle phenomena specific to Chinese.

In the presence of English sentences annotated for negation scope and their translations in a target language, projection via word alignment information has a low recall, showing that most negation instance are rendered as positive construction or lost in the alignment noise. We found annotation projection not to yield good results (F1 <0.5 on avg.) for any of the negation components we have annotated – cue, scope and event. We have also shown that the relatively lower recall means that effort should be put in recovering those negation instances that projection have failed to identify, which is less desirable than correcting incorrect projections. Finally, we also show that projection errors are mostly due to the fact that negation instances can be translated into positive constructions, as well as due to alignment errors.

Negation scope can be detected across languages on top of Universal Dependencies parses where syntactic annotations are not language-specific. In Chapter 5, we used NEGPAR to assess whether a model trained in a language can be used in another where no annotated data is available. To bridge the gap between these two languages we leverage two intermediate representations: Universal Dependencies (UD), a syntactic annotation framework consistent across language, and cross-lingual word embeddings.

Negation scope detection is done on top of UD parses, under the assumption that certain dependency substructures around and within the scope of negation are represented consistently in both English and Chinese and that lexical information is provided by the cross-lingual embeddings.

Structured neural models can be used for cross-lingual negation scope detection, with recursive architectures yielding the best performance. However, BiLSTMs are not suitable for the task since they overfit to the training data. We experimented with two different neural architectures that can take as input UD parses: a bidirectional Dependency LSTM (biDLSTM), which is an extension to a treeLSTM of Tai et al. (2015), and a Graph Convolutional Network (Marcheggiani and Titov, 2017, GCN). We compare the performance of these model against the BiLSTM model in both a monolingual and in a cross-lingual setting.

We found that it is indeed possible to build a model transferable from a source to a target language, although performance is still worse compared to a monolingual oracle system. We also found the BiDLSTM model to always outperform the GCN model. On the other hand, a sequence classifier is not a good fit for cross-lingual negation scope detection since it overfits to the training data.

All models are guided in their prediction mostly by structural information and less by word-embedding information. Through a feature ablation experiment we have shown that word embeddings are not an important features in negation scope detection; **structure is what matters in negation scope detection, both within and across languages**. Recurrent models use punctuation to guide prediction, while structural models use information from the input parse.

Structured models performs differently depending on the dependency substructure the scope is in and they still fail in capturing some of the phenomena related to negation scope. Through an error analysis we have shown that our best structured models tend to perform better when the cue is the same substructure as its scope. This concerns in particular cases where the scope spans the whole sentence, a clausal complement or a coordinated clause. We also found that phenomena requiring lexical information, such as neg. raising, as still entirely missed by the model.

6.2 Moving forward: Universal Dependencies to logical form with negation scope

The focus of this thesis has been detecting *string-level* negation scope across languages. While doing so, one of our main findings concerns the role of Universal Dependencies: a cross-lingual model relies mostly on the parse structure and on the dependency labels to detect the span of text in the scope of negation, with little contribution from lexical information in the form of cross-lingual word embeddings.

In future work, we would to expand these conclusions further by asking: *is it possible to leverage Universal Dependencies to represent negation scope at the level of logical form as well?*

If one could indeed rely mostly on the labels and the tree structure to represent the scope of negation, and given that the annotations are common across all UD treebanks, one could then consider representing the formal semantics of negation scope in multiple languages. This will also address the problem that most semantic banks are limited to English (e.g. GMB and 'DeepBank' Flickinger et al. (2012)) or only to a few other languages (e.g. 'The Spanish Resource Grammar' Marimon (2010)).

In preliminary work, we attempted to convert the annotations in the UD to logical form using a preexisting rule-based framework, *UDepLambda* (Reddy et al., 2017, 2016), that we extended to deal with negation scope phenomena. The framework first assigns each dependency label a manually crafted lambda expression that represents its meaning (e.g. $nusbj \rightarrow \lambda f.\lambda P \lambda Q.P(\lambda e.Q(\lambda y.Agent(e,y))$, for further detail on the semantics used we refer the reader to Fancellu et al. (2017)) and then reduces these using the semantics of the head and governor words. Reduction is performed by traversing the tree in a manually specified order.

However, tailoring these lambda expressions to represent the meaning of each of the dependency labels soon becomes a challenging task. The same dependency label might in fact correspond to different roles (and therefore lambda expressions), depending on the context; whereas the relationship holding between the words 'I' and 'cook' in the sentence 'I cook pizza' can be defined as 'Agent', in 'I hate you' the subject is the 'Experiencer' of the event 'hate'.

For this and other reasons, a more robust alternative to rule-based methods is to learn a probabilistic mapping between a UD tree and its correspondent logical form. We address future work in this direction.

We envision this system to be an encoder-decoder architecture with attention that takes as input a UD parse and outputs a directed acyclic graph representing the correspondent logical form.

There are two aspects worth considering.

One concerns the *data*; we need in fact a semantic bank where negation scope, among other semantic phenomena is fully specified. We choose this to be the Parallel Meaning Bank Abzianidze et al. (2017), where semantics is annotated on top of raw text in four different languages –English, Italian, Dutch and German– using Discourse Representation Theory(Kamp et al., 2011), which is FOL translatable. This is important considering that the long-term goal is to represent negation scope as a logical form in multiple languages and that we require data in languages other than English for evaluation (similar to the role that NEGPAR had in this thesis).

The other aspect concerns the *model* itself.

The architecture that best fits our problem would be one where the encoder encodes a tree and the decoder reconstructs a directed acyclic graph. Encoding the input UD tree is straightforward given that one could use a treeLSTM or one of its variants, as seen in Ch. 5. On the other hand, decoding the semantic graph is not.

A solution worth investigating would be to build the target semantic graph incrementally as a series of actions. The system learns *jointly* a) which fragments should make up the entire graph, b) how they should be composed together and c) in which order. Future work may take into consideration recent work on graph decoding (Li et al., 2018) or on graph grammars (Groschwitz et al., 2015) as a starting point.

Appendix A

Notation

ACC: accusative (case)

ADV: adverbial

ASP: aspectual particle

BA: Chinese light verb 把, introducing the direct object in pre-verbal position.

CL: classifier

DE: Chinese particle 的, introducing relative clauses.

EMPH: Particles indicating emphasis

GEN: genitive

GUO: 过, aspectual particles indicating an action that has already been experienced

IMP: imperative

IND: indicative (mood)

INTER: interrogative particle

NE: sentence-final question particle 呢, used when the subject has already mentioned by the speaker. NEG: negation marker

PAST: past (tense)

PAST.PL: past participle (tense)

PART: particle

PASS: passive

PRES: present (tense)

PROGR: progressive aspect

SING: singular

TOP: topic marker

1PERS: first person

Appendix B

NEGPAR: annotation guidelines

B.1 Cue

Negation cues are characters or words inherently expressing negation. There are a total of 45 negation cues in Chinese, most of which are adverbs. Amongst these we can identify 10 *core* negation cues, which are one-character words (except for 没有“did/have not”) that can be combined with other words to form compound cues (see § B.1.3). For instance, the cue 没有in (96.i) can be compounded with the character 并to reinforce its negative meaning as shown in (96.ii) (in the examples throughout this document the cue is marked in **bold**, the scope is underlined once and the event inside a box).

(96) i 遗憾 的 是 咱们 **没有** 遇到 他
pity DE be we did not meet him
It is a pity we did not meet him

ii 我 原 以为 我 摆弄 手杖 的 事 **并没有** 叫 他 发觉 呢 。
I originally think I play cane DE thing never let him realize NE .
‘Originally I thought he would never find out I was playing with the cane’

Certain cues can also function as morphemes and be affixed to adjectives; for instance, the most common cue in Chinese, the adverb 不 (roughly equivalent to the English ‘not’) can be used both as a stand-alone word and as affixal negation (不贵, literally ‘inexpensive’).

B.1.1 Core negation cues in Chinese

B.1.1.1 不

不 is the most common cue in Chinese and has similar distributional properties to the English ‘not’, except the it cannot negate existentiality. As shown in (97), 不 is often used to mark verbal negation in the present.

(97) 我 不 知道 应该 相信 什么 。

I not know should believe what .

“I don’t know what I should believe.”

B.1.1.2 没(有) and 未

没(有) is a negated auxiliary verb that marks an even that has not been completed or achieved. Unlike 不 it cannot be used with habitual events or events in the present. Sometimes the character 有 can be omitted. An example is shown in (98).

(98) 我 没(有) 看见 他

I did-not see he

“I did not see him.”

未 is the classical form of 没(有) and has the same meaning of “have not/did not” as shown in (99).

(99) 一 件 尚 未 得到 解释 的 事实

a CL yet have-not get explanation DE fact

“A fact that has not been explained yet.”

B.1.1.3 没and 无

The cue 没 is used to negate both existentiality (‘there is/are’) and possession (‘have’), which in Chinese are expressed by the same verb, 有. Although homographs, this is different from the cue 没(有) marking negation on a past event, where the character 有 does not bear either the meaning of ‘there is’ or ‘have’. An example of 没 as negating possession is shown in (100).

(100) 我 没 有 理由

I not have reason

“I have no reasons.” [11.69]

无 is the literary form of the existential 没有, ‘there is not/have not’. In modern Chinese, 无 is also used as a negation prefix equivalent to English “-less”. e.g. 无线(“no wire=wireless”).

B.1.1.4 别, 勿and 莫

Unlike English, where mood is encoded in verbs and auxiliaries only, Chinese can place mood information on the cue itself. This is the case of the imperative which has a dedicated cue 别, along with its classical forms 勿 and 莫. (101) exemplifies its use.

(101) **别** 这样， 华生
not.IMP this , Watson
“Don’t be like this, Waston” [4.180]

B.1.1.5 非

非is the literary form of 不是, “is not”. 非 is commonly used with the adverb 并, as a compound cue 并非, as shown in (102).

(102) 死亡 **并非** 由于 自然 原因
death not due to natural causes
His death was not due to natural causes

B.1.1.6 否

否is used as negation prefix in front of verbs, like in 否认(“not-acknowledge = deny”).

B.1.2 Affixal negation in Chinese

Affixal negation is problematic to define in Chinese, as it is difficult to define what a morpheme is and if Chinese even exhibits morphology. For instance, whereas in ‘inexpensive’ the negation morpheme ‘in-’ in English is bound to the word and cannot exist independently, in Chinese its translation ’不贵’ can be decomposed into 不, ’not’ and 贵, ’expensive’, which possess status as individual words.

Whichever status we assign to such forms in Chinese, they should be annotated in Chinese alongside their scope because unlike in English they create contradictions and not contraries. For instance where in English ‘It is neither expensive nor inexpensive’ is a valid statement where the object in question is somewhat expensive, constructions where an adjective and its negation are false at the same time are not allowed in Chinese.

Just for convenience and to contrast these construction against the core negation cues used in isolation, we will use the term ’affixal negation’.

Except 别 and 勿, all the other core negation cues can function as a prefix in a compound such as 没用(无用– “useless”). Some of them are modern prefixes that were created through translation, e.g. 无线(“no wire=wireless”), 非物质(“not physical=non-physical”), 不道德(“not moral=immoral”).

Unlike English, the cue 不 in Chinese can be infixed. This is the case of resultative and potential constructions that are realized by a verb + cue + complement, where the latter can indicate direction or result. This is exemplified in (103), where in the verb compound 听到, literally “to hear-arrive”, the second character marks the result of the main verb “to hear”. In these constructions, the negative cue placed in between signifying that the result cannot be achieved. (As a matter of fact, a more intuitive interpretation of infixal negation in Chinese is that the result introduced in the complement is ‘unachievable’).

(103) 我 听不到
I hear-not-arrive
“I could not hear”

B.1.3 Compound cues

Chinese also exhibits constructions where the cue marker is preceded by an adverb expressing a degree of emphasis. This involves, amongst others, the compounds 并不(“not”),绝不(“absolutely not”),决不(“absolutely not”) and 绝无(“absolutely no”), where 并, 绝, 决are all bound morphemes with 并only occurring alongside negation. Given that none of the previous morphemes can stand independently, we annotate them as part of the cue. (104) exemplifies this decision:

(104) 他 决不 是 我 在 这里 所 见到 过 的 人
He not be I in here all seen GUO DE man
‘He is no one whom I have seen down here’

B.1.4 Discontinuous cue

Some constructions allows for the same cue to be discontinuous. This is the case of the Chinese construction 既不...也不, equivalent to English cue “neither...nor”, exemplified in (105).

(105) 对 他 既不 应该 可怜 , 也不 应该 原谅
towards him not should pity , not should excuse
[...] for whom there was neither pity nor excuse

It is worth mentioning that in such constructions, the adverbs 既 and 也 can be omitted as in the case of the expression 不骄不躁(既不骄也不躁, "neither proud nor upset").

Exceptions in Chinese are flagged by the discontinuous cue 除了...之外, equivalent to the English "save for", "except".

(106) 他 除了 时常 彻夜 不 眠 之外 ,
早晨 总是 起 得 很晚 的
he except often all night not sleep except ,
morning always wake up ADV very late DE
[...] was usually very late in the morning, save upon those not infrequent occasions when he was up all night.

B.1.5 False negation cues

Both in English and Chinese, some negative affixes do not introduce negative polarity and should not be annotated as cue. For instance, the word "disgrace" in English does not mean "a lack of grace" and the negative meaning carried by "dis", as well as its status as morpheme, only exists etymologically. Although this is less of a problem in Chinese, where most compounds containing affixal negation are semantically transparent, we still abide by the following three criteria to determine the presence of a false negation affix.

Compositionality. When the meaning of a compound can be derived from its parts, we mark the cue as such. For example, 无聊("boring") in Chinese is not equivalent to the meaning of its parts, "no chatting", hence we do not annotate 无 as a cue. In the same way, we do not annotate the 不 in 不见("to have disappeared") where the meaning of the two characters "not" and "appear" combined do not yield the meaning of the whole word.

Obsolescence. When the meaning of the morpheme modified by the negative affix is obsolete, the affix should not be annotated as the negation cue. For example, 然 in 不然("otherwise") and 则 in 否则("otherwise") have the meaning of "like this" in classical Chinese, but not in Modern Chinese.

Non-negative meaning. If the compound has a positive meaning, the cue should not be annotated. This is the case of words that introduce emphatic degree modification such as 无比(lit. “not compare” - “very”), 不已(lit. “not stopped” - “very much”) which all introduce false negative affixes.

For the same reason, besides cases of false affixal negation, we also do not annotate cues in rhetorical questions, where the speaker uses negation to confirm rather than to deny a statement. This includes constructions such as 你不觉得...? (‘Don’t you think...?’), as shown in (107), or 这不正是？(‘Isn’t it exactly...’).

(107) 你 不 觉得 很 有 趣味 吗 ？
you not think very have interest INTER ?
Don’t you find it interesting?

There are also several expressions such as 毫无疑问(“without a doubt”), 无疑(“no doubt”) where the speaker expresses his certainty regarding a statement or a fact using a negative construction. In cases like (108) we do not annotate the cue.

(108) [...] 无疑 地 早 就 会 拨转 马头 回去 了
no-doubt ADV early then can turn horse-head return ASP
[...] would have been right glad to have turned his horse’s head

Similarly there are cases of discontinuous false negation cues. For example, 不(“not”)是(“is”)... 就(“then”)是(“is”) is always translated as “either...or”. Another similar construction is 不是别人(“not others) ... 正是(“is exactly”) which can be translated into the English “none other than...”. An example of these constructions is shown in (109).

(109) 不 是 他 就 是 我 总 有 一 个 得 穿上 捆疯子 用 的 紧身衣 的
not be he then be I always there-is one CL must wear insane use DE straitjacket DE
‘Either I or him ought to be in a straitjacket’

Negation cues in fixed pragmatic expressions are also not annotated as such since the overall meaning is not negative. This concerns in particular expressions such as 对不起(“not able to treat you well - sorry”) and 没关系(“not a matter - It’s all right. “).

Similar to English, we do not annotate negation in the expressions 不能不and 不得不, ”cannot help but”, both having a reinforced positive meaning of “must, have to”. Same goes for the discontinuous cue 非得...不可. (110) exemplifies this.

(110) 我 不得不 放弃 这 种 方法
I must abandon this CL method
I am compelled to abandon this method.

We do not annotate the negation cue in idioms if these have positive meanings. This is the case of (111), where the idiom 无往(而)不利 contains an instance of double negation yielding an overall positive meaning ("there are no places where victory is not achieved = always successful").

(111) 你 应该 就 把 这 事 也 记下来 , 作为 我 无往而不利 的 反证 吧
you should then BA this thing also record , do I always-successful DE disproof ASP
[...] if you are an honest man you will record this also and set it against my successes!

Finally, we do not annotate the cue in the expression 说不定(lit. 'not able to say definitely' – 'maybe') which express possibility as shown in (112).

(112) 他们 说不定 信任 我们
They maybe trust us
'They maybe trust us'

B.2 Scope

We define negation scope as the sentence span affected by the presence of the negation cue. We consider here a semantic notion of negation scope: in general, if negation directly affects an event, the scope should also include its argument and modifiers as shown in 113.

(113) 咱们 **没有** 遇到 他
we have-not meet him
'We haven't met him'

The scope can also be discontinuous. This includes cases of long-range dependencies where material is omitted from the negated clause but can be retrieved from other spans of the sentence; as shown in (114), this is often the case of coordinated clauses where the object is only referenced in the first clause.

(114) 我 把 他 弃 而 不 顾 了
I BA he abandon and not care ASP
I abandoned and not cared about him.

In English, to determine whether a sentence span is in the scope of negation, the "it is not the case that" test is often used. This involves paraphrasing the negated sentence as positive preceded by the expression “it is not the case that” and checking whether they have the same meaning (e.g. ‘I don’t eat pizza’ and its paraphrase ‘It is not the case that I eat pizza’). In Chinese, one can use the correspondent Chinese phrase ‘并不是...’ to test for negation scope, as shown in (115):

(115) 我 不 吃 比萨
I not eat pizza
I don’t eat pizza
并不是 我 吃 比萨
not-be I eat pizza
It is not the case that I eat pizza

B.2.1 Coordinate clauses

In coordinate clauses, negation scope spans only the clause containing the cue. (114) exemplifies this, where only the verb 顾, ‘to care’, is negated whereas ‘弃’, ‘to abandon’ is not. If any argument is omitted but can retrieved from other spans of the sentence, this is also included in the scope, as in the case of the subject 我,‘I’ and 把他, ‘him’ in (114).

B.2.2 Subordinate clauses

If negation appears in a subordinate clause, we include the subordinate and not the matrix clause in the scope of negation. This is exemplified in (116).

(116) 我 本 想 找 借口 不 听 他 说
I originally think find excuse not listen he say
‘I was originally thinking of finding excuses not to listen to him.’

Here the cue ‘不’ denies the event in the infinitival, ‘听他说’ (‘listening to him’), which is included in the scope, but not the event in the matrix clause, ‘找借口’ (‘finding excuses’) which is excluded from it.

On the other hand, if negation appears in the matrix clause, we exclude the subordinate from its scope, as shown in (117).

(117) 在 我 对 此 事 作出 决定 之前 ， 什么 也不 告诉 他
in I towards this thing do decision before , anything not tell he
‘And then you will not say anything to him until I make up my mind on that matter
.’

The only exception to this are conditional statements, such the one introduced by ‘要是’ and ‘如果’ where both the matrix and the subordinate are in the scope of negation when negation appears in the former. This is shown in (118).

(118) 如果能 对 你 有 帮助的 话 ， 我就 不 出去 了
if can towards you have help DE speak , I then not go out ASP
‘If I can receive any help with you, I won’t go out’

B.2.3 Sentence final particles

In Chinese, there are a number of Sentence-final modal particles such as 吗, 呢, 呀, 哇whose purpose is to express the speaker’s attitude towards an utterance. We do not annotate these particles in the scope of negation. Our decision is also supported by previous work that defines these sentence-final particles complementisers out of IP (Paul, 2014). For example in (119), the sentence-final particle 呀 expresses emphasis towards a statement but doesn’t contribute to the overall meaning of the sentence.

(119) 不 要 等 他过 了 山 呀!
not need wait he pass ASP mountain ASP
‘There is no need to wait until he has passed the mountain!’

Notice that some of these particles are multifunctional. For instance, sentence-final 的 can function both as a marker for the genitive, as in (120), as well as to reinforce the meaning of a sentence as in (121).

(120) 这 是我的 钱包 ， 不 是你 的
this is I DE wallet , not is you DE
‘This is my wallet, not yours.’

(121) 我是不 肯 帮 那些 坏蛋 的
I be not surely help those rascal EMPH .
“I refused to help those rascals! “

In (120) the particle 的 is used as a marker for genitive, indicating possession; similar to English, Chinese allows the possessed item, ‘wallet’, to be omitted if already mentioned in the preceding clause. In (121) instead 的 is used to strengthen the meaning of the sentence, with a usage roughly corresponding to an exclamation mark in English. Scope is annotated only in cases like (120) but not in ones like (121).

Finally, we always annotate the sentence-final aspectual particle 了as part of the scope, since it specifies the aspect of the negated event as in (122).

(122) 我 问 不 出 什么 东西 了
I ask not out any thing ASP
"I cannot ask anything (from her) any more."

B.2.4 Non-sentential negation on subject

Chinese does not allow constructions where negation is directly expressed on an indefinite pronoun such as the case of "nothing", "nobody" and "nowhere" in English. Instead, these are rendered in Chinese with the negated existential 没有("do not exist"). As shown in (123) when the subject is negated the entire clause falls under the scope of negation.

(123) 没 有 人 注意到 它们
not exist person notice them
"No one noticed them."

B.2.5 Negated adjectives

The scope of a negated adjective is the adjective itself and does not include the noun it specifies. It is to be noted that the negation does not scope over coordinate adjectives in the same noun phrase. When more than one adjective specifies a noun, as in (124), the negation annotation does not scope over the coordinate adjective 废弃 as it is not affected by the negation cue.

(124) 那 人 住 在 这些 废弃 不用 的 小房 中
That person lives in these deserted not-used DE little.house inside
"That person lives in these deserted and unused little houses. "

B.2.6 Negated adverbs

When it is an adverb to be directly negated, the scope of should span only the adverb itself and not the event it modifies. This is exemplified in (125).

(125) 心 里 感到 不安 地 驾车
heart in feel insecure ADV drove
"Drove while feeling insecure."

In Chinese entire clauses can also function as adverbial modifiers on the main verb. As shown in (126), these constructions usually take the form of verb+得+adverbial clause. Again, if negation is in the adverbial clause only this is included in the scope of negation, along with any argument that can be recovered from other parts of the sentence.

(126) 那 牧 人 当时 被 吓 得 简直 都说 不 出 话 来
that CL man at that time PASS fear ADV simply all speak feel not go out speak come * ASP
“and the man, was so crazed with fear that he could barely speak.”

B.2.7 Relative Clauses

If a cue negated an event inside a relative clause, we consider this as its scope. Unlike English, in Chinese relative clauses precede the noun they modify, with the marker 的 usually placed in between. This is shown in (127).

(127) 他 是 个 不 爱 出风头 的 人 。
he is CL not like show-off DE person .
“He is not a person who likes to show off.”

Chinese also allows the noun to be omitted in cases where it has been mentioned in a preceding part of the sentence, with negation scope still scoping on the relative clause only. This is exemplified in (128) where the noun 人, ”man”, becomes the subject of the sentence.

(128) 那 个 人 是 不 爱 出风头 的
that CL man is not like show-off DE
“He is not a person who likes to show off.”

B.2.8 Exceptions

Certain cues in Chinese introduce exceptions. This is the case of the discontinuous cue “除了...之外”; its use is exemplified in (129).

(129) 他 除了 时常 彻夜 不 眠 之外 ,
早晨 总是 起 得 很晚 的
he **except** often all night not sleep **except** ,
morning always wake up ADV very late DE
[...] was usually very late in the morning, save upon those not infrequent occasions when he was up all night.

Here the exception isolates those instances where the event of “waking up very late” does not apply and is therefore negated. We define these cases as the “exception to positive” and annotate the span of text in between the cue “除了...之外” as part of the cue.

The case in (129) contrasts with the “exception to nothing” where the cue is used to exclude a set of instances for which negation does not apply. This is shown in (130)

(130) 除了 帮助 他 之外 ， 没有 其他 目的
Except help he except , there are no other purposes
”He doesn’t have any purposes other than helping him.”

In this case, the material in between the discontinuous cue specifies those instances that are positive, i.e. for which the subject had a purpose. In cases like this we do not annotate the construction “除了...之外” as a cue, and as a consequence we do not annotate a scope either.

B.2.9 Comparative constructions

In Chinese, comparison is expressed in most cases through the co-verb ‘比’ , which takes as subject and object the two things compared, followed by the dimension they are compared along. This is the case in (131), where the subject and the object are compared for their age; in cases like this, we annotate as scope the entire clause.

(131) 约翰森 先生 年纪 不 比 你 他
Johnson Mr. age not compare you old
‘Mr. Johnson is not older than you.’

However, negation can also exclude this dimension. We distinguish these cases from the one in (131), by excluding from the scope in the object of the comparison.

B.2.10 Lexical items marking universal quantification

Lexical items marking universal quantification are excluded from the scope of negation when they scope over it. In Chinese, this concerns mostly the adverb 都and the adjectives 全部and 所有(all corresponding to the English ”all”), as shown in (132).

(132) 我 不 把 我 知道 的 全部 事情 都 说出来
I **not** BA I know DE all thing all speak
I do not speak about things I do not know about

Notice that 都can also be used in certain constructions without the meaning of ”all” in the English guideline is “even”. For instance, in the construction 甚至连...都, corresponding to the English ”even”, 都does not mark the presence of universal quantification and is therefore included in the scope. This is shown in the following example.

(133) 我 甚至 连 帽子 都 没有 戴
I even even hat all not wear
”I couldn’t even wear a hat”

In Chinese cases where negation scopes over universal quantification (similar to the English ”not every”) are realized by negation preceding the adjective 所有. In this case the lexical items expressing quantification is included in the scope, as shown below.

(134) 不是 所有 的 人 来 参加 我 的 晚会
Not all DE people come participate I DE party
”Not everyone will come to my party”

B.2.11 Neg raising

In cases of neg raising – the phenomenon by which certain negated predicates (e.g. ”think”, ”believe”, ”expect”) can give rise to a reading where the negation seems to take scope from an embedded clause, we annotate the embedded clause only but not the main clause in the scope of negation. This is shown in (135).

(135) 我 真 想不到 会 看见 这样 长长 的 头颅
I very think-not-arrive could see this king very long DE
I never thought I could see such a long skull

In the example above, despite negation marking directly the event 想到, ”to think”, it is the object of the thought to be negated and scope is therefore interpreted on the subordinate only.

B.2.12 Modality

We use the Chinese counterpart of the “It is not the case...“ test, ”并不是...”, to identify whether a modal scopes inside negation or viceversa.

Most modals signaling epistemic modality are annotated inside the scope of negation. This includes modals such as 应该(”would”), 会, 能, 可以(”can/could”), etc. The example below shows an example annotation.

(136) 这样 的 一 个 人 就 不 会 迁往 乡村 去 了
this kind DE one CL person then not drift country go ASP
”such a person would not drift into the country.”

Both in English and in Chinese the example above can be paraphrased into ”it is not the case that (such as person would drift into the country)”, with the modal ”would” inside the scope of negation.

On the other hand, some deontic modals are excluded from the scope of negation. These include most imperative constructions that in Chinese can be realized by the modal 要 following the negation cue 不, as shown in (137).

(137) 不 要 动
not must move
”Don’t move”

Considering that a literal translation of the example in (137) is ”You must not move”, this is not equal to ”It is not the case you must move”, since the original sentence does not allow for a meaning where ”you don’t have to move”.

Notice that, depending on the surrounding context, certain instances of 要have the meaning of ”have to”; in these we annotate it in the scope of negation. This is shown in (138), where the Chinese sentence can be translated literally into ”You don’t have to move”, which has the same meaning of ”It is not the case that you have to move”.

(138) 你 不 要 动 ; 你 可以 站在 这里
you not have to move ; you can stand still here
”You don’t have to move; you can stand still here”

B.2.13 Interrogative pronouns

Finally, if a negated clause includes any interrogative pronouns we include these in the scope of negation. This is exemplified by the pronoun 为什么(“why”) in (139):

(139) 为什么 不 到 房子 里面 呢
why not go to house inside Q.PRT?
“Why did you not go into the house?”

B.3 Event

We annotate an event as negated if it is factual; the term ‘factuality’ includes here both states and nominal elements. What the annotation considers as an event is a minimal unit in a negated phrase, usually corresponding to its head. An example of annotation of a verbal predicate negated event is shown in (140), where the event is presented inside a $\boxed{\text{box}}$ (we omit the scope just for presentational purposes). Although one could consider 吃羊肉, ‘eat mutton’, as the entire event, the event is just its minimal unit, that is, the head verb 吃 ‘to eat’

(140) 我 不 $\boxed{\text{吃}}$ 羊肉
I not eat mutton
‘I do not eat mutton.’

The event of a negated verb phrase should be the main verb. Therefore we do not annotate any aspectual markers as part of the event. For example, in the following example, the durative aspect marker 在 is excluded from the annotation of the event that only spans on the verb 动, ”to move”.

(141) 没 有 人 在 $\boxed{\text{动}}$
Not exist one PROGR move
“Nobody is moving”

Notice that sometimes the aspectual marker occurs within a verb phrase. For example in (142), the experiential marker 过 marker should be excluded from the annotation of the event 结婚, ”to get married”.

(142) 他 还 没有 $\boxed{\text{结}}$过$\boxed{\text{婚}}$ 啊
He still not married-GUO-married ASP
“He is still not a married man.”

B.3.1 Copular constuctions

In the case of copular constructions we annotate as event the head of the NP in the predicate. For example in (143), only the head of the negated predicate, 朋友(“friend”), is marked as the event.

(143) 他 的 朋友 也 不 是 我 的 $\boxed{\text{朋友}}$
he DE friend also not be I DE friends
“His friends are not also mine.”

B.3.2 Adjectival predicates

Unlike English, where adjectives can appear in the main predicate only following a copula, in Chinese adjective form predicates without a copula. The following example contrasts the one in (143) where the adjective marked as an event is not preceded by the copula 是.

(144) 这样 不 公平
this way not fair
"This is not fair."

B.3.3 Existential constructions

We do not annotate as an event the existential verb 有("there is/are") when negated. Instead, we mark the head of the following noun phrase as the event. This is shown in (145).

(145) 没有 希望
there is not hope
'There is no hope'

When existentiality is marked on the subject (with a meaning similar to the determiner "no" in English), we mark the verb head in the predicate as the event. This is the case of 住, "lived", in (146).

(146) 没有 人 住 过 这
there are no person line EXP here .
"No one lived here."

B.3.4 Identifying Non-Factuality

We do not mark as events those appearing in the following non-factual constructions.

Imperatives. We have already mentioned how in Chinese imperatives are often introduced by deontic modals which are excluded from the negation scope.

For those negated imperatives marked by the cue 别, we do not annotate the event because orders and requests are not factual, as shown below.

(147) 别 动
not move
"Don't move"

Non-factual interrogatives. Most non-rhetorical yes-no questions are non-factual. In (148) for instance, the speakers asks for confirmation about a negative statement that was uttered previously; however, the event of ”being present” has not been confirmed and it is therefore considered non-factual.

(148) 那 位 女士 不 在 吗 ?
that CL lady not present INTER ?
“The lady is not there?”

On the other hand, we assume that questions introduced by interrogative pronouns are factual. For example in (149), the event of “not going to the house” is factual because it happened. Therefore, 到(“go to”) is marked as the event.

(149) 为什么 不 $\boxed{到}$ 房子 里面 去 拜访 呢 ?
why not arrive house inside go visit NE ?
“Why did (you) not go into the house to visit (him)?”

Conditional constructions. Given that we do not annotate hypothetical events, we also do not mark events in both a conditional clause and the main clause containing a conditional subordinate.

Modality. Modality is typically related to non-factuality as it mostly expresses possibility and necessity. Therefore we usually do not annotate events in clauses introduced by modal verbs or verbs that express modality.

However, modal verbs expressing participant internal ability are annotated as events. The Chinese equivalent of “can”, 能, is also multifunctional and can introduce either ability or possibility. We should mark 能as an event when it means “able to” as in (150). This also applies to other modal verbs that can express participant-internal ability such as 能够, 会,可以and 可.

(150) 我 不 能 早一点 到 那里 去
I not can earlier arrive there go .
“I can’t go there earlier.”

Supposition and presumption. To make a judgment on the factuality, one should also examine the semantics of the verb that introduces the scope. If this suggests the speaker’s certainty about the content of the following clause (e.g. 确信(“to be sure”), 确定(“to be certain”), 知道(“to know”)), we should treat the embedded event as factual. If, however, the verb suggests that the following statement is part of the speaker’s supposition or presumption as in the case of 相信(“to believe”), 认为(“to believe”), 觉得(“to think”), 想(“to think”), 害怕(“to fear”) etc., the negated event in the statement should not be marked as shown in (151).

(151) 我 相信 您 **决不** 愿意 做 一 个 妨碍 别人 的 人 。
I believe you not want do one CL hinder others DE person .
“I believe that you do not wish to be a spoil-sport.”

Future tense. As a language with no morphological tense marker, Chinese employs various linguistic devices to indicate future tense. This is the case of temporal expressions such as 明天(“tomorrow”), connectives such as 以后(“later on”) or adverbs such as 将(“going to”). Given that the event has not happened yet and is therefore not factual, we do not annotate it as such. This is exemplified in (152).

(152) 我 以后 **再也不** 提 这 件 事 了 。
I later on never mention this CL thing ASP .
“I will not mention this matter again.”

B.3.5 Verb-complement compounds

Chinese allows for complex verbal structures where a main verb is followed by a complement indicating result or potentiality, as described in the case of infixal cues in § B.1.2.

When annotating the event we distinguish two cases: if the meaning of the whole construction can not be derived compositionally, we annotate the entire construction as the event. For example, the following verb-complement compounds will be annotated as a whole when being negated: 看见 (“see-appear: see”), 听到 (“hear-arrive: hear”), 看出来 (“see-out: see”), 弄明白 (“make clear: understand”). Otherwise, the event is just the main verb as in 说不清楚(“speak-not-clearly:cannot speak clearly”).

B.3.6 Idioms as negated event

Idiomatic phrases in Chinese are generally made up of four characters. When negated, these expressions as a whole are treated as an event. An example is shown in (153), where the idiom 引以为耻 is the main predicate of the sentence.

(153) 我 **并不** 把 和 他 妹妹 的 感情 引以为耻 。
I not BA with he sister DE feelings ashamed .
“I am not ashamed of my feelings towards his sister.”

Notice however, if the negation cue is found within the set phrase and if the idiom is semantically transparent – i.e. its overall meaning can be composed by the meaning of the single characters, we will identify a specific element in the phrase as the event. For example, the negated event in the idiom, 无可估量(“not able to estimate”), should be 可(“can”) alone.

Bibliography

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., Van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *arXiv preprint arXiv:1702.03964*.

Altuna, B., Minard, A.-L., and Speranza, M. (2017). The scope and focus of negation: A complete annotation framework for italian. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 34–42.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Baker, C. L. (1970). Double negatives. *Linguistic inquiry*, 1(2):169–186.

Baker, K., Bloodgood, M., Dorr, B. J., Callison-Burch, C., Filardo, N. W., Piatko, C. D., Levin, L. S., and Miller, S. (2012). Modality and Negation in SIMT Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, 38(2):411–438.

Ballesteros, M., Díaz, A., Francisco, V., Gervás, P., de Albornoz, J. C., and Plaza, L. (2012). UCM-2: a Rule-Based Approach to Infer the Scope of Negation via Dependency Parsing. In Agirre, E., Bos, J., and Diab, M. T., editors, **SEM@NAACL-HLT*, pages 288–293. Association for Computational Linguistics.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). UGroningen: Negation detection with Discourse Representation Structures. In Agirre, E., Bos, J., and Diab, M. T., editors, **SEM@NAACL-HLT*, pages 301–309. Association for Computational Linguistics.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Bergsma, S. and Van Durme, B. (2011). Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1764.

Blanco, E. and Moldovan, D. I. (2011). Some Issues on Detecting Negation from Text. In Murray, R. C. and McCarthy, P. M., editors, *FLAIRS Conference*. AAAI Press.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017a). The groningen meaning bank. In *Handbook of Linguistic Annotation*, pages 463–496. Springer.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017b). *The Groningen Meaning Bank*, pages 463–496. Springer Netherlands, Dordrecht.

Calixto, I., Liu, Q., and Campbell, N. (2017). Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*.

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics.

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

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Chen, H., Huang, S., Chiang, D., and Chen, J. (2017). Improved neural machine translation with a syntax-aware encoder and decoder. *arXiv preprint arXiv:1707.05436*.

Cohen, S. B., Das, D., and Smith, N. A. (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 50–61. Association for Computational Linguistics.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2-3):281–332.

Councill, I. G., McDonald, R., and Velikovich, L. (2010). What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 51–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cruz, N. P., Taboada, M., and Mitkov, R. (2016). A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

Dahl, Ö. (1979). Typology of sentence negation. *Linguistics*, 17(1-2):79–106.

de Albornoz, J. C., Plaza, L., Díaz, A., and Ballesteros, M. (2012). UCM-I: A Rule-based Syntactic Approach for Resolving the Scope of Negation. In Agirre, E., Bos, J., and Diab, M. T., editors, **SEM@NAACL-HLT*, pages 282–287. Association for Computational Linguistics.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–92.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262. Association for Computational Linguistics.

Dinu, G., Lazaridou, A., and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Dryer, M. S. (2005). Negative morphemes. *The world atlas of language structures*, pages 454–457.

Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 845–850.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Fancellu, F., Reddy, S., Lopez, A., and Webber, B. (2017). Universal dependencies to logical form with negation scope. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 22–32.

Fancellu, F. and Webber, B. (2015). *Translating Negation: A Manual Error Analysis*, pages 2–11. Association for Computational Linguistics.

Fancellu, F. and Webber, B. L. (2014). Applying the semantics of negation to SMT through n-best list re-ranking. In Bouma, G. and Parmentier, Y., editors, *EACL*, pages 598–606. The Association for Computer Linguistics.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Feagin, C. (1979). Negation. In *Variation and change in Alabama English: A sociolinguistic study of the White community*, pages 209–242. Georgetown University Press, Washington, DC.

Flickinger, D. (1999). The english resource grammar.

Flickinger, D., Zhang, Y., and Kordoni, V. (2012). Deepbank. a dynamically annotated treebank of the wall street journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.

Forest, R. (1993). *Négations: essai de syntaxe et de typologie linguistique*, volume 77. Peeters Publishers.

Gella, S., Sennrich, R., Keller, F., and Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Groschwitz, J., Koller, A., and Teichmann, C. (2015). Graph parsing with s-graph grammars. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1481–1490.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2016). A representation learning framework for multi-source transfer parsing. In *AAAI*, pages 2734–2740.

Haspelmath, M. (2001). *Indefinite pronouns*. OUP Oxford.

Haspelmath, M. (2005). Negative indefinite pronouns and predicate negation. *The world atlas of language structures*, pages 466–469.

Hermann, K. M. and Blunsom, P. (2013). Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 58–68.

Horn, L. (1989). A natural history of negation.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392–399. Association for Computational Linguistics.

Jackendoff, R. S. (1969). An interpretive theory of negation. *Foundations of language*, pages 218–241.

Jia, L., Yu, C., and Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1827–1830. ACM.

Kamp, H., Van Genabith, J., and Reyle, U. (2011). Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer.

Khemlani, S., Orenes, I., and Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Klima, E. S. (1964). *Negation in english*. na.

Konstantinova, N., de Sousa, S. C. M., Díaz, N. P. C., López, M. J. M. n., Taboada, M., and Mitkov, R. (2012). A review corpus annotated for negation, speculation and their scope. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *LREC*, pages 3190–3195. European Language Resources Association (ELRA).

Labov, W. (1972). Negative attraction and negative concord in english grammar. *Language*, pages 773–818.

Ladusaw, W. (1979). Negative polarity items as inherent scope relations. *Unpublished Ph. D. Dissertation, University of Texas at Austin*.

Ladusaw, W. A. (1980). Polarity sensitivity as inherent scope relations.

Lapponi, E., Velldal, E., Øvrelid, L., and Read, J. (2012). UiO 2: Sequence-labeling Negation Using Dependency Features. In Agirre, E., Bos, J., and Diab, M. T., editors, **SEM@NAACL-HLT*, pages 319–327. Association for Computational Linguistics.

Lauly, S., Boulanger, A., and Larochelle, H. (2014). Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Le, P. and Zuidema, W. (2014). The inside-outside recursive neural network model for dependency parsing. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 729–739.

Li, C. N. and Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Li, J., Zhou, G., Wang, H., and Zhu, Q. (2010). Learning the Scope of Negation via Shallow Semantic Parsing. In Huang, C.-R. and Jurafsky, D., editors, *COLING*, pages 671–679. Tsinghua University Press.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Marcheggiani, D. and Titov, I. (2017). Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Marimon, M. (2010). The spanish resource grammar. In *LREC*.

Miestamo, M. (2007). Negation–an overview of typological research. *Language and Linguistics Compass*, 1(5):552–570.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Morante, R. and Blanco, E. (2012). * sem 2012 shared task: Resolving the scope and focus of negation. 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*, pages 265–274. Association for Computational Linguistics.

Morante, R. and Daelemans, W. (2009). A Metalearning Approach to Processing the Scope of Negation. In Stevenson, S. and Carreras, X., editors, *CoNLL*, pages 21–29. ACL.

Morante, R. and Daelemans, W. (2012). ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *LREC*, pages 1563–1568. European Language Resources Association (ELRA).

Morante, R., Liekens, A. M. L., and Daelemans, W. (2008). Learning the Scope of Negation in Biomedical Texts. In *EMNLP*, pages 715–724. ACL.

Morante, R., Schrauwen, S., and Daelemans, W. (2011). Annotation of negation cues and their scope: Guidelines v1. *Computational linguistics and psycholinguistics technical report series, CTRS-003*.

Mrkšić, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., and Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association of Computational Linguistics*, 5(1):309–324.

Mutalik, P. G., Deshpande, A., and Nadkarni, P. M. (2001). Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents. *Journal of the American Medical Informatics Association*, 8:598–609.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., et al. (2017). Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

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: Long Papers)*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Paul, W. (2014). Why particles are not particular: Sentence-final particles in chinese as heads of a split cp. *Studia Linguistica*, 68(1):77–115.

Payne, J. R. (1985). Negation. language typology and syntactic description. vol. 1 (clause structure), ed. by t. shopen, 197–242.

Postolache, O., Cristea, D., and Orasan, C. (2006). Transferring coreference chains through word alignment. In *Proceedings of LREC-2006*.

Pražák, O. and Konopik, M. (2017). Cross-lingual srl based upon universal dependencies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 592–600.

Pullum, G. K. and Huddleston, R. D. (2002). Negation.

Qian, Z., Li, P., Zhu, Q., Zhou, G., Luo, Z., and Luo, W. (2016). Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 815–825.

Read, J., Velldal, E., Øvrelid, L., and Oepen, S. (2012). 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*, pages 310–318. Association for Computational Linguistics.

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., and Lapata, M. (2016). Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

Reddy, S., Täckström, O., Petrov, S., Steedman, M., and Lapata, M. (2017). Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.

Reitan, J., Faret, J., Gambäck, B., and Bungum, L. (2015). Negation Scope Detection for Twitter Sentiment Analysis. In Balahur, A., van der Goot, E., Vossen, P., and Montoyo, A., editors, *WASSA@EMNLP*, pages 99–108. The Association for Computer Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *LREC*.

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859.

Søgaard, A., Agić, Ž., Alonso, H. M., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual nlp. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487. Association for Computational Linguistics.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Tiedemann, J. (2015). Cross-lingual dependency parsing with universal dependencies and predicted pos labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Tsuruoka, Y. and Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 467–474. Association for Computational Linguistics.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292:247.

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.

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(S-11).

Vincze, V., Szarvas, G., Móra, G., Ohta, T., and Farkas, R. (2011). Linguistic scope-based and biological event-based speculation and negation annotations in the bioscope and genia event corpora. *Journal of Biomedical Semantics*, 2(5):S8.

Vulić, I. and Moens, M.-F. (2013). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116.

Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Wetzel, D. and Bond, F. (2012). Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing. In Carpuat, M., Specia, L., and Wu, D., editors, *SSST@ACL*, pages 20–29. Association for Computational Linguistics.

White, J. P. (2012). Uwashington: Negation resolution using machine learning methods. 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*, pages 335–339. Association for Computational Linguistics.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao, M. and Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Zhang, Y. and Barzilay, R. (2015). Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In *LREC*.

Zou, B., Zhou, G., and Zhu, Q. (2013). Tree kernel-based negation and speculation scope detection with structured syntactic parse features. In *EMNLP*, pages 968–976.

Zou, B., Zhou, G., and Zhu, Q. (2016). Research on Chinese negation and speculation: corpus annotation and identification. *Frontiers of Computer Science*, 10(6):1039–1051.