SICK-NL: A Dataset for Dutch Natural Language Inference

Gijs Wijnholds
UiL-OTS
Utrecht University
g.j.wijnholds@uu.nl

Michael Moortgat
UiL-OTS
Utrecht University
m.j.moortgat@uu.nl

Abstract

We present SICK-NL (read: *signal*), a dataset targeting Natural Language Inference in Dutch. SICK-NL is obtained by translating the SICK dataset of Marelli et al. (2014) from English into Dutch. Having a parallel inference dataset allows us to compare both monolingual and multilingual NLP models for English and Dutch on the two tasks. In the paper, we motivate and detail the translation process, perform a baseline evaluation on both the original SICK dataset and its Dutch incarnation SICK-NL, taking inspiration from Dutch skipgram embeddings and contextualised embedding models. In addition, we encapsulate two phenomena encountered in the translation to formulate *stress tests* and verify how well the Dutch models capture syntactic restructurings that do not affect semantics. Our main finding is all models perform worse on SICK-NL than on SICK, indicating that the Dutch dataset is more challenging than the English original. Results on the stress tests show that models don't fully capture word order freedom in Dutch, warranting future systematic studies.

1 Introduction

One of the primary tasks for Natural Language Processing (NLP) systems is Natural Language Inference (NLI), where the goal is to determine, for a given premise sentence whether it contradicts, entails, or is neutral with respect to a given hypothesis sentence.

For English, several standard NLI datasets exist, such as SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). Having such inference datasets available only for English may introduces a bias in NLP research. Conneau et al. (2018) introduce XNLI, a multilingual version of a fragment of the SNLI dataset, that contains pairs for Natural Language Inference in 15 languages and is explicitly intended to serve as a resource for evaluating crosslingual representations. However, Dutch is not represented in any current NLI dataset, a lack that we wish to complement.

Dutch counts as a high-resource language, with the sixth largest Wikipedia (2M+ articles), despite having ca. 25M native speakers. Moreover, the syntactically parsed LASSY corpus of written Dutch (van Noord et al., 2013), and the SONAR corpus of written Dutch (Oostdijk et al., 2013) provide rich resources on which NLP systems may be developed. Indeed, Dutch is in the scope of the multilingual BERT models published by Google (Devlin et al., 2019), and two monolingual Dutch BERT models have been published as part of HuggingFace's transformers library (de Vries et al., 2019; Delobelle et al., 2020).

Compared to English, however, the number of evaluation tasks for Dutch is limited. There is a Named Entity Recognition task coming from the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003); from a one million word hand annotated subcorpus of SONAR (Oostdijk et al., 2013) one derives part-of-speech tagging, Named Entity Recognition and Semantic Role Labelling tasks. More recently a Sentiment Analysis dataset was introduced, based on Dutch Book reviews (van der Burgh and Verberne, 2019). Moreover, Allein et al. (2020) introduce a classification task where a model needs to distinguish between the pronouns *die* and *dat*.

Given the focus on word/token-level tasks in Dutch, we aim to complement existing resources with an NLI task for Dutch. We do so by deriving it from the English SICK dataset, for the following reasons: first, this dataset requires a small amount of world knowledge as it was derived mainly from image captions that are typically concrete descriptions of a scene. Therefore, no world knowledge requirements will be imposed on an NLP model

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1474–1479
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

for the task, but rather its ability for reasoning will be assessed. Secondly, due to the structure of the sentences in SICK, the types of inferences can be attributed to particular constructs, such as hypernymy/hyponymy, negation, or choice of quantification. Thirdly, SICK contains 6076 unique sentences and almost 10K inference pairs, making it a sizeable dataset for NLP standards, while deriving a Dutch version is more manageable than with other datasets. We make the dataset, code and derived resources (see Section 5) available online[1].

2 Dataset Creation

We follow a semi-automatic translation procedure to create SICK-NL, similar to the Portuguese version of SICK (Real et al., 2018). First, we use a machine translator to translate all of the (6076) unique sentences of SICK[2]. We review each sentence and its translation in parallel, correcting any mistakes made by the machine translator, and maintaining consistency of individual words' translation, in the process guaranteeing that the meaning of each sentence is preserved as much as possible. Finally, we perform a postprocessing step in which we ensure unique translations for unique sentences (alignment) with as few exceptions as possible. In this way we obtain 6059 unique Dutch sentences, which means that the dataset is almost fully aligned on the sentence level. It should be noted, however, that we can not fully guarantee the same choice of words in each sentence *pair* in the original dataset, as we translate sentence by sentence. In the whole process, we adapted 1833 of the 6076 automatic translations, either because of translation errors or alignment constraints. As we are interested in collecting a comparable and aligned dataset, we maintain the relatedness and entailment scores from the original dataset.

Table 1 shows some statistics of SICK and its Dutch translation. The most notable difference is that the amount of unique words in Dutch is 23% higher than that in English, even though the total number of words in SICK-NL is about 93% of that in SICK. We argue that this is due to morphological complexities of Dutch, where verbs can be separable or compound, leading them to be split up into multiple parts (for example, "*storing*" becomes "*opbergen*", which may be used as "*de man bergt iets op*"). Moreover, Dutch enjoys a relatively free word order, which in the case of SICK-NL means that sometimes the order of the main verb and its direct object may be swapped in the sentence, especially when the present continuous form ("*is cutting an onion*") is preserved in Dutch ("*is een ui aan het snijden*"). Finally, we follow the machine translation, only making changes in the case of grammatical errors, lexical choice inconsistencies, and changes in meaning. This freedom leads to a decrease in relative word overlap between premise and hypothesis sentence, computed as the number of words in common divided by the length of the shortest sentence. From the perspective of Natural Language Inference this is preferable as word overlap often can be exploited by neural network architectures (McCoy et al., 2019).

	SICK	**SICK-NL**
No. of tokens	189783	176509
No. of unique tokens	2328	2870
Avg. sentence length	9.64	8.97
Avg. word overlap	66.91%	58.99%

Table 1: Basic statistics of SICK and SICK-NL.

3 Baseline Evaluation and Results

We evaluate two types of models as a baseline to compare SICK-NL with its English original.

First, we evaluate embeddings that were not specifically trained on SICK. Table 2 shows the correlation results on the relatedness task of SICK and SICK-NL, where the cosine similarity between two independently computed sentence embeddings is correlated with the relatedness scores of human annotators (between 1 and 5).

SICK		**SICK-NL**	
Skipgram	69.49	Skipgram	56.94
BERT$_{cls}$	50.78	BERTje$_{cls}$	49.06
BERT$_{avg}$	61.36	BERTje$_{avg}$	55.55
RoBERTa$_{cls}$	46.62	RobBERT$_{cls}$	43.93
RoBERTa$_{avg}$	62.71	RobBERT$_{avg}$	52.33

Table 2: Pearson r correlation coefficient for the relatedness task of the English SICK dataset (left) and its Dutch translation (right).

To obtain sentence embeddings here, we average skipgram embeddings, or, in the case of contextualised embeddings, we take either the sentence embedding given by the $[CLS]$ token, or

[1] github.com/gijswijnholds/sick_nl

[2] We used DeepL's online translator www.deepl.com/translator

we take the average of the individual word's embeddings. For the skipgram embeddings in English, we use the standard 300-dimensional GoogleNews vectors provided by the `word2vec` package and for Dutch, we use the 320-dimensional Wikipedia trained embeddings of Tulkens et al. (2016).

The relatedness results show that (a) using the $[CLS]$ token embedding as a sentence encoding performs worse than taking the average of word embeddings, and that (b) the Dutch incarnation of SICK is harder than the original English dataset. It may be noted that relatedness scores are less robust then entailment labels, and so our first result may not be enough support for the claim that SICK-NL poses a more challenging task. For example, it could be that relatedness scores will differ slightly if we were to ask a number of annotators to re-evaluate the Dutch dataset.

In the second setup, we use BERTje, the Dutch BERT model of de Vries et al. (2019) and RobBERT, the Dutch RoBERTa model of Delobelle et al. (2020), with their corresponding English counterparts, as well as multilingual BERT (mBERT), as sequence classifiers on the Entailment task of SICK(-NL). Here we observe a similar pattern in the results in Table 3: while there are individual difference on the same task, the main surprise is that the Dutch dataset is harder, even when exactly the same model (mBERT) is used.

	SICK		**SICK-NL**
BERT	87.34	BERTje	83.94
mBERT	87.02	mBERT	84.53
RoBERTa	90.11	RobBERT	82.02

Table 3: Accuracy results on the entailment task of the English SICK dataset and its Dutch translation for two Dutch BERT models and their English counterparts. For each model, we report the best score out of 20 epochs of fine-tuning.

4 Error Analysis

In order to understand the differences between the Dutch and English language models on the respective tasks, we dive deeper into the classification results. We plot confusion matrices for each model in Table 5, where we separate predictions that the models have in common and the from the predictions that are unique to each model.

In the case of English, performance on classifying contradictions is worse for multilingual BERT and RoBERTa, and RoBERTa also gives highest recall values for the Neutral and Entailment labels. This is all not surprising given that RoBERTa has the overall highest test set accuracy. The surprising results come mainly from the comparison between English and Dutch models. Where BERTje is rather indecisive when it comes to Neutral sentence pairs (it classifies roughly equal numbers as Neutral and Entailment), it classifies 74% of Entailment pairs as Neutral. For multilingual BERT the situation is reversed, with 47% of Neutral entailments classified as Entailment, although for cases of entailment, the classifier did not clearly distinguish Neutral from Entailment. The most surprising pattern was observed in RobBERT: where RoBERTa still has high recall for Neutral and Entailment, its Dutch counterpart RobBERT mistakes most Neutral cases as Entailment and even more so vice versa. For all models, in these four cases of misclassification, in the case of the English task the correct inference was made in at least 99% of the cases.

Following Naik et al. (2018), we inspect these prominent cases of misclassification in Dutch by looking at the number of cases of high overlap (at most four words not in common), and at the number of length mismatches (the difference between sentence length exceeds 4), and set off these distributions against that of the test set, in Table 4.

	BERT (N→E)	**mBERT** (E→N)	**RobBERT** (N→E)	**RobBERT** (E→N)	**Test**
			Word difference		
EN	66%	42%	62%	44%	40%
NL	47%	38%	41%	38%	28%
			Length mismatch		
EN	24 %	17%	25%	17%	27%
NL	30 %	31%	28%	25%	36%

Table 4: Error analysis of prominent misclassifications.

The main finding here is that word overlap does provide a strong cue in the English dataset, especially given that SICK has more cases (1970) overall than SICK-NL (1385), and that in SICK they are more concentrated in cases of Entailment. Length mismatches occur more often in SICK-NL but seem to provide less of a cue to the models to make strong inference decisions.

BERT		Prediction EN-NL				Prediction EN				Prediction NL			
		C	N	E	*rec.*	C	N	E	*rec.*	C	N	E	*rec.*
Gold	C	549	62	16	88%	46	27	12	54%	22	52	11	26%
	N	35	2341	147	93%	37	147	83	55%	32	116	119	43%
	E	3	143	1035	88%	3	53	167	75%	3	165	55	25%
	pr.	94%	92%	86%		53%	65%	64%		39%	35%	30%	
mBERT													
Gold	C	553	69	10	88%	24	47	9	30%	42	31	7	52%
	N	32	2344	106	94%	9	210	89	68%	71	93	144	30%
	E	3	160	1015	86%	0	103	123	54%	15	111	100	44%
	pr.	94%	91%	90%		73%	58%	56%		33%	40%	40%	
RoBERTa													
Gold	C	563	68	4	89%	27	46	4	35%	35	28	14	45%
	N	25	2301	82	96%	6	279	97	73%	79	96	207	25%
	E	1	108	995	90%	2	42	256	85%	15	246	39	13%
	pr.	96%	93%	92%		77%	76%	72%		27%	26%	15%	

Table 5: Confusion matrices for English vs Dutch language models, finetuned. Top: BERT vs BERTje. The models disagree in 13.3% of cases. Middle: Multilingual BERT. The model disagrees in 14.3% of cases). Bottom: Roberta vs RobBERT. The models (disagree in 18.3% of cases).

5 Stress Testing

One of the potential sources of error could have been the passive form translation of a verb. Such constructions, combined with a prepositional phrase, form an interesting testbed for Dutch as they allow the prepositional phrase to be moved in front of the verb in a sentence without changing the meaning. For example, "*Een vrouw is aan het wakeboarden op een meer*" ("*A woman is wakeboarding on a lake*"), may in Dutch be used interchangeably with "*Een vrouw is op een meer aan het wakeboarden*"). We select all (87) sentences in SICK-NL that contain both the 'aan het' construction and a prepositional phrase, and generate their permutations. Then, we replace all (225) inference pairs with these sentences such that they now contain a sentence with different word order but the exact same meaning and therefore the inference label is preserved. We then verify how the model's predictions do on those inference pairs that were in the test set (116). Additionally, we check whether the models are able to interchange sentences and their rewritten equivalent (i.e. classify as Entailment).

As a second test, we investigate the role of the simple present versus the present continuous. We take all the (383) cases of present continuous in the Dutch dataset and replace them by a simple present equivalent, leading to 1137 pairs, out of which 576 occur in the test data. For example, we turn the sentence "*De man is aan het zwemmen*" into the simple form "*De man zwemt*". We then repeat the same procedure as above, asking how many inference predictions change as a result of this form change, and whether the forms can be used interchangeably for the models.

	present cont. → present simple			
	Before	After	→	←
BERT	84.55	86.63	93.21	92.43
mBERT	86.11	84.90	94.26	94.52
RobBERT	82.81	81.94	86.16	84.33
	prep. phrase order switch			
BERT	81.03	78.45	85.06	85.06
mBERT	87.93	85.34	85.06	80.46
RobBERT	76.72	75.86	72.41	73.56

Table 6: Stress test accuracy. Left: accuracy before and after rewriting. Right: inference between rewritings.

The results in Table 6 indicate that the interchange between present continuous and simple present forms does not make much of a difference to the models' performance, and interchangeability is high except for RobBERT that scores under 90%. However, switching the order of prepositional phrase and verb has a much stronger effect with all models consistently scoring lower on the relevant part of the test set, and mainly the models being particularly poor at interchanging these sentences that are semantically equivalent.

6 Conclusion

In this paper we introduced an NLI dataset for Dutch by semi-automatically translating the SICK dataset. To our knowledge this is the first available inference task for Dutch. Despite the common perception that Dutch is very similar to English, SICK-NL was significantly more difficult to tackle, even for language models that had access to the training data for fine-tuning. We hypothesised that the difference in result may be due to a larger vocabulary in SICK-NL, and a decline in word overlap between inference pairs. In addition we performed two stress tests and found that pretrained models that were exposed to the training data had difficulty detecting semantically equivalent sentences that differ only in word order. Further work will therefore more systematically assess such phenomena.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable feedback, which improved the paper. Both authors acknowledge support from the Dutch Research Council (NWO) under the scope of the project "A composition calculus for vector-based semantic modelling with a localization for Dutch" (360-89-070).

References

Liesbeth Allein, Artuur Leeuwenberg, and Marie-Francine Moens. 2020. Binary and multitask classification model for Dutch anaphora resolution: Die/dat prediction. *arXiv preprint arXiv:2001.02943*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Benjamin van der Burgh and Suzan Verberne. 2019. The merits of universal language model fine-tuning for small datasets–a case with Dutch book reviews. *arXiv preprint arXiv:1910.00896*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based language model. *arXiv preprint arXiv:2001.06286*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 216–223, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. *The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch*, pages 219–247. Springer Berlin Heidelberg, Berlin, Heidelberg.

Livy Real, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruna Thalenberg, Bruno Guide, Cindy Silva, Guilherme de Oliveira Lima, Igor C. S. Câmara, Miloš Stanojević, Rodrigo Souza, and Valeria de Paiva. 2018. SICK-BR: A Portuguese corpus for inference. In *Computational Processing of the Portuguese Language*, pages 303–312, Cham. Springer International Publishing.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural*

Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4130–4136, Portorož, Slovenia. European Language Resources Association (ELRA).

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.