Probing for Dutch Relative Pronoun Choice

Author1 AUTHOR1@EMAIL.COM

Affiliation, Address, Country

Abstract

We propose a linguistically motivated version of the relative pronoun probing task for Dutch (where a model has to predict whether a masked token is either *die* or *dat*), collect realistic data for it using a parsed corpus, and probe the performance of four context-sensitive BERT-based neural language models. The task turns out to be much harder than the original version, which simply masked all occurrences of *die* and *dat*. Models differ considerably in their performance, but a monolingual model trained on a heterogeneous corpus appears to be most robust.

1. Introduction

Delobelle et al. (2020) test RobBERT, a BERT-based neural language model for Dutch, on a masked language prediction task where the model has to predict for a given sentence whether the masked token is either *die* or *dat*. Probing tasks like this have been used to investigate to what extent neural language models are sensitive to linguistic structure and linguistic concepts such as the distinction between singular and plural subjects, case or gender (see Linzen and Baroni (2020) for an overview). In contrast to statistical ngram language models and 'static' neural language models such as word2vec (Mikolov et al. 2013), which take only a limited context into account, context-sensitive neural language models of the BERT-variety (Devlin et al. 2019) have access to all positions in the input to make a prediction about the masked position. The task is inspired by a similar experiment from Allein et al. (2020). Allein *et al.* use data from Europarl and Sonar to train a neural classifier for *die/dat*-prediction, using an LSTM that is initialized with word embeddings obtained from a word2vec model. The trained classifier obtains accuracies of 83.2% (Europarl) and 84.5% (Sonar).

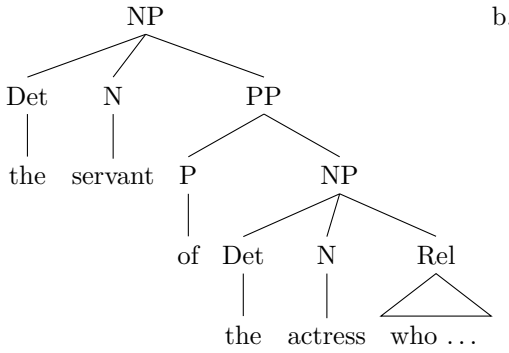
Delobelle et al. collect data from the Dutch section of the Europarl corpus, where they use all sentences containing *die* or *dat* as test cases (288K sentences). They report accuracies of 90.2 (mBERT), 94.9 (Bertje) and 98.7% (RobBERT). It should be noted, though, that in many cases, there is no real ambiguity or need to pay attention to longer contexts, as (1) *dat* is used as subordinating conjunction in a position where *die* could never occur, (2) *die* or *dat* is used but there are no preceding nouns with the opposite gender that could function as distractor, or *die* or (3) *dat* is used as deictic pronoun introducing an NP (*die jongen, that boy*), where the noun with which the pronoun agrees is usually adjacent or very close to the pronoun (without intervening distractors). As a probing task, this dataset is therefore less suitable, as it is unclear to what extent information from the full context of the masked position is required to make the correct prediction.

We propose to turn the *die/dat* prediction task into a proper probing task by focussing on cases where a relative pronoun occurs in the masked position (thus ignoring prediction of the relatively easy conjunction and deictic pronoun cases) and where there is at least one distractor. We explain how we collected relevant examples from a parsed corpus, and test two monolingual and two multilingual neural language models for their ability to make the right predictions and thus establish to what extent such models are sensitive to longer contexts.

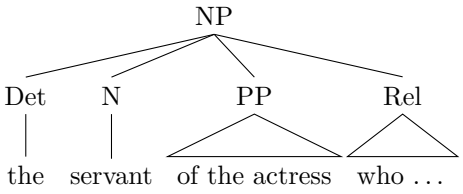
2. Relative Clause Attachment

Relative clause attachment has been studied extensively from a psycholinguistic perspective (starting with Cuetos and Mitchell (1988)), also for Dutch (Desmet et al. 2006). The canonical example is *Someone shot the servant of the actress who was on the balcony.* Here, the relative clause could be attached either to the lower noun *actress* as in (1a) or the higher noun *servant* as in (1b). It has been claimed that preference for high or low attachment is not universal, with speakers of some languages having a preference for low attachment, where other languages, such as Dutch, prefer high attachment (Brysbaert and Mitchell 1996). Desmet et al. point out that corpus data shows that in Dutch, low attachment is the most frequent configuration as well, and high attachment only is more frequent in cases where the highest noun is animate. The preference of test subjects is in line with this observation if one takes these factors into account during construction of the test items.

(1) a.



b.



Whereas psycholinguistic studies have used a sentence completion task to probe the preference for test subjects for either high or low attachment of the relative clause, we change the task into a masked word prediction task suitable for testing a language model by only masking the relative pronoun. As the language model is trained on raw text and never sees any syntactic structures, we exclusively concentrate on those cases where the two preceding nouns in the complex noun phrase have opposite gender (*neuter* or *non-neuter*, respectively), thus triggering either *die* or *dat*, which we take as signs of high or low attachment in this particular context.

(2) a. Het is het proza$_{neu}$ van een vrouw$_{nonn}$ [MASK] een hoge prijs betaalde
It is the writing of a woman [MASK] paid a high price

b. Melk is het enige product$_{neu}$ van de koe$_{nonn}$ [MASK] aan de paniek is ontsnapt
Milk is the only product of the cow [MASK] has escaped from the panic

(2a) illustrates low attachment, where the relative clause is modifying the lower noun *vrouw*, and the correct pronoun is *die*$_{nonn}$. (2b) illustrates high attachment, where the relative clause modifies *product*, and the masked pronoun therefore is *dat*$_{neu}$.

3. Data Collection

To collect realistic examples, we searched a newspaper corpus, containing articles from *Algemeen Dagblad* and *NRC Handelsblad* 1994-1995. This corpus has not been used during training of the language models, thus there cannot be memorization effects (Carlini et al. 2020). The corpus was automatically parsed with Alpino (van Noord 2006). Our corpus query searched for sentences containing a complex NP with the structure of either (1a) or (1b) and with the additional constraint that the two relevant nouns had to be singular and of opposite gender. Note that we did not impose

Model	Corpus	Training objective	Reference
RobBERT	Dutch section of OSCAR (web, 39GB)	MLM	(Delobelle et al. 2020)
Bertje	Dutch Wikipedia, Sonar, Novels (12GB)	MLM	(de Vries et al. 2019)
mBERT	Wikipedia for 104 languages	MLM and NSP	Github[a]
XLM-R	Common Crawl (web, multilingual, 2.5TB)	MLM	(Conneau et al. 2020)

Table 1: Neural-language models used in the probe.

a. https://github.com/google-research/bert/blob/master/multilingual.md

any other constraints. In particular, any number of other modifying elements (adjectives, other PPs) can be present, and the relative clause may be extraposed.

The high attachment instances returned by the search query contain a substantial number of false hits (due to parsing mistakes), so we manually selected 1951 true positives from the 2534 hits returned by the query. Results for the low-attachment query hardly contain false hits, so we did not filter these. As low attachment is also much more frequent than high attachment and we wanted to use balanced data, we included 2000 instances of low-attachment. Note that while false hits (and duplicate sentences) were removed, sentences that actually are ambiguous after masking the pronoun were kept. An example is:

(3) Een dagboek van de nieuwe NCMV-topman [MASK] een tip oplicht van de geheimen van de onderhandeling .
A diary of the new NCMV-head [MASK] reveals the secrets of the negotiation

Here, the choice for *dat* amounts to choosing *dagboek* as antecedent (in line with the source) whereas *die* amounts to choosing *NCMV-topman* as antecedent, which is a semantically plausible interpretation as well.

4. Language Models

We tested on two multilingual and two monolingual context-sensitive language models (Table 1). Whereas statistical (ngram) language models are typically trained to make a prediction about the next token in a sentence, neural language models are trained using a masked-language model (MLM) objective, where some words in the text are masked and the model has access to both left and right context to make a prediction. All models of the BERT-variety (Devlin et al. 2019) use a complex neural architecture that includes an attention mechanism that allows the model to learn which words in the input are most important for making a prediction about the masked position. The question thus arises whether this enables the model to learn that, say, the head noun inside a complex subject NP is the relevant word for predicting the correctly inflected form of the finite verb, or whether the model instead learns a simpler but structure insensitive rule (such as attending to the closest preceding noun) that would allow it to make the right prediction in most cases without paying attention to structure. Some models also use a next-sentence prediction (NSP) task for training, where the model is trained to predict whether two concatenated sentences occurred in the given order in the corpus or not. While this can be relevant for downstream tasks such as question answering, it is probably less relevant for learning to attend to linguistic structure. A language model can be trained on a monolingual corpus, but recently it has been shown that multilingual models, trained on a concatenation of monolingual corpora (with upsampling of the data for low-resource languages), can outperform monolingual models especially for low-resource languages (Pires et al. 2019). We used the pretrained models present on huggingface.co.[1]

1. https://huggingface.co/GroNLP/bert-base-dutch-cased, https://huggingface.co/pdelobelle/robbert-v2-dutch-base, https://huggingface.co/xlm-roberta-base, https://huggingface.co/bert-base-multilingual-cased

	High attachment					Low attachment				
	N	Bertje	RobBERT	mBERT	XLM-R	N	Bertje	RobBERT	mBERT	XLM-R
dat	788	**0.789**	0.537	0.510	0.590	1000	**0.931**	0.782	0.769	0.852
die	1163	0.740	0.650	0.711	**0.783**	1000	0.937	0.836	0.885	**0.957**
overall	1951	**0.761**	0.601	0.624	0.700	2000	**0.934**	0.809	0.827	0.905

Table 2: Probing results for high and low attachment

5. Experiments and Discussion

The results of the probing experiment are in Table 2. Note that in order to do well on this task, a model must take both left and right context into account. The left context provides two nouns, while the right context contains the relative clause. In order to decide on the correct pronoun, the model must work out whether the relative clause is more likely to be attached to the higher or lower noun in the left context. The high attachment cases are much harder than low attachment for all models. This is not unexpected, as low attachment is more frequent in the corpus and probably in the training corpora as well, and high attachment requires the model to attend to a noun that is relatively far from the masked position with an intervening noun that has the opposite gender. Bertje is the only model that does equally well on the *die* and *dat* cases, where the other models all have a tendency to prefer *die* over *dat*. Apart from Bertje, the multilingual model XML-R suffers least from the difference between *die* and *dat* and therefore does better than the other two models.

With a monolingual model doing best and worst on the task, it cannot be concluded that monolingual models are always to be preferred over multilingual models (or vice versa). A more important factor might be the corpus used for training the language models. The probing sentences are all relatively complex, due to the fact that we select sentences with a complex NP containing a PP and a relative clause. Such sentences might be more frequent in corpora that inlcude newspaper text than in corpora consisting of text from Wikipedia or the web only. In the newspaper corpus we used, we estimate that around 80% of the relevant cases are instances of low attachment. It might well be that in other corpora the distribution is even more skewed, making it increasingly hard for the model to learn to make the right predictions for high attachment cases. The non-neuter relative pronoun *die* is somewhat more frequent than *dat* (approx. 40 and 60%, respectively) in our data. It seems that this cannot explain the tendency for some models to prefer *die*, although this preference might be stronger in more recent corpora, especially if they also contain informal text (Audring 2013, Bouma 2017). The fact that we can test for high or low attachment by simply masking the pronoun is due to the fact that Dutch relative pronouns agree with the antecedent. In light of the discussion from the psycholinguistics literature on (speakers of) languages having a preference for either high or low attachment, it would be interesting to repeat this probe for other languages. For languages that do not have relative pronoun agreement, one might use singular and plural nouns, and relative clauses where these have the subject role, so the verb can be masked to test whether the model can predict the correctly inflected form.

References

Allein, Liesbeth, Artuur Leeuwenberg, and Marie-Francine Moens (2020), Automatically correcting Dutch pronouns 'die' and 'dat', *Computational Linguistics in the Netherlands Journal* **10**, pp. 19–36. https://www.clinjournal.org/clinj/article/view/102.

Audring, Jenny (2013), A pronominal view of gender agreement, *Language Sciences* **35**, pp. 32–46. https://www.sciencedirect.com/science/article/pii/S0388000112001088.

Bouma, Gosse (2017), Agreement mismatches in Dutch relatives, *Belgian Journal of Linguistics* **31** (1), pp. 136–163, John Benjamins.

Brysbaert, Marc and Don C Mitchell (1996), Modifier attachment in sentence parsing: Evidence from Dutch, *The Quarterly Journal of Experimental Psychology Section A* **49** (3), pp. 664–695, SAGE Publications Sage UK: London, England.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. (2020), Extracting training data from large language models, *arXiv preprint arXiv:2012.07805*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020), Unsupervised cross-lingual representation learning at scale, *Proceedings of the ACL*, pp. 8440–8451.

Cuetos, Fernando and Don C Mitchell (1988), Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish, *Cognition* **30** (1), pp. 73–105, Elsevier.

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

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: a Dutch RoBERTa-based Language Model, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, pp. 3255–3265. https://www.aclweb.org/anthology/2020.findings-emnlp.292.

Desmet, Timothy, Constantijn De Baecke, Denis Drieghe, Marc Brysbaert, and Wietske Vonk (2006), Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account, *Language and Cognitive Processes* **21** (4), pp. 453–485, Taylor & Francis.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://www.aclweb.org/anthology/N19-1423.

Linzen, Tal and Marco Baroni (2020), Syntactic structure from deep learning, *Annual Review of Linguistics*, Annual Reviews.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), Distributed representations of words and phrases and their compositionality, *NIPS*, pp. 3111–3119.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019), How multilingual is multilingual BERT?, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 4996–5001. https://www.aclweb.org/anthology/P19-1493.

van Noord, Gertjan (2006), At last parsing is now operational, *in* Mertens, Piet, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42.