SOPalign: A Tool for Automatic Estimation of Compliance with Medical Guidelines

Luke van Leijenhorst[1,3] [0000−0002−3929−1285], Arjen P. de Vries[1] [0000−0002−2888−4202], Thera Habben Jansen[2] [0000−0003−3064−4774], and Heiman Wertheim[3] [0000−0002−5003−5565]

[1] Institute for Computing and Information Sciences, Radboud University, Nijmegen, Netherlands
{arjen.devries, luke.vanleijenhorst}@ru.nl
[2] Department of Infection Prevention and Control, Amphia Hospital, Breda, Netherlands
thabbenjansen@amphia.nl
[3] Department of Medical Microbiology, Radboudumc, Nijmegen, Netherlands
heiman.wertheim@radboudumc.nl

Abstract. SOPalign is a tool designed for hospitals and other healthcare providers in the Netherlands to automatically estimate the compliance of internal standard operating procedures (SOPs) for employees with the national guidelines. In this tool, users can upload the SOPs of their hospital and the recommendations from the most recent guidelines. SOPalign will then link the individual recommendations from the guidelines to the relevant passages of text in the SOPs and determine whether these passages are compliant with the recommendations. To link the SOP passages to the recommendations from the guideline, we make use of a Semantic Textual Similarity (STS) model based on the siamese BERT-network architecture. For efficiency reasons, we only apply the STS model to sentences that exceed a threshold in n-gram cosine similarity. To estimate compliance of SOPs with guideline recommendations, we have fine-tuned pre-trained language models using two different Dutch Natural Language Inference (NLI) datasets.

Keywords: Information Retrieval · Natural Language Inference · Semantic Textual Similarity.

1 Introduction

In healthcare, medical knowledge is captured into guidelines that support healthcare professionals to deliver the best possible quality of care. In hospitals and other healthcare settings, these guidelines are implemented and integrated in local *standard operating procedures (SOPs)*. When guidelines are updated, it is important to also update all SOPs that are based on this guideline. In most hospitals, this is a manual process in which employees periodically adjust the SOPs associated with the guideline in the quality system. Given the number of

guidelines and the number of SOPs in a hospital, this is a vulnerable process that may lead to SOPs that are (partially) outdated.

SOPalign is a tool that uses natural language processing techniques to assess where SOPs are compliant with the applicable guidelines, and where they differ. The tool can support healthcare facilities to keep their procedures updated with the guidelines. It also helps to discover bottlenecks in the guideline recommendations, when recommendations in guidelines do not appear in SOPs, or the SOPs are not compliant with the guideline. This is an important step in the improvement cycle of guidelines and supports the perspective of 'living guidelines'. The authors foresee a transition from a collection of independent documents into a hyperlinked network of guidelines and SOPs. Our tool aims to enable this transition at a low cost.

2 Tool design

An overview of SOPalign can be found in Fig. 1. The input consists of a file with a list of recommendations drawn from the guidelines we want to validate, and the PDFs of the SOPs we want to analyse. Before we analyse the SOPs, we extract the text fully automatically from the PDFs and attempt to merge these lines of text into their original paragraphs. These passages of text, together with the recommendations from the guidelines are then passed to the guidelines linker.

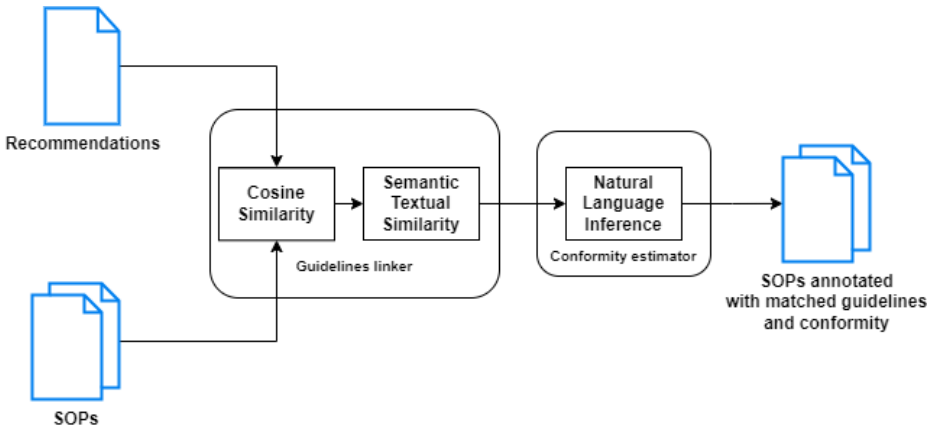

Fig. 1. Architecture of the SOPalign tool.

Guidelines linker. The goal of the guidelines linker is to link every recommendation of the guidelines to their corresponding passages of text in the SOPs to identify where this recommendation is implemented. Because a guideline can

contain up to a hundred recommendations, and every recommendation gets coupled to every passage of every SOP, this generates a very large candidate set. Before we use our large Semantic Textual Similarity (STS) language model to assess whether passages correspond to the recommendations from the guideline, we calculate the character n-gram based cosine similarity between the SOP passage and the guideline recommendations, and only keep pairs above a certain threshold so the candidate set becomes substantially smaller. By doing so, we make the assumption that at least some character n-grams of the recommendation will also be in the SOP passage where it is implemented.

Now we feed the smaller dataset to our multilingual STS language model. We make use of the *paraphrase xlm-r-multilingual-v1* model which is constructed from sentence embeddings using siamese networks [1]. We only keep the instances where this STS score exceeds a certain threshold. All the SOP passage and recommendation pairs that exceeded both the cosine similarity and STS thresholds are linked together and are sent to the Compliance estimator. In our application of SOPalign, for a set of 329 SOPs and 104 guidelines, there were roughly 5.200.000 candidate pairs. Of this number, 7611 ($\approx 0.15\%$) exceeded the cosine threshold, and of those pairs 543 ($\approx 7.13\%$) exceeded both thresholds.

Compliance estimator. The compliance estimator determines whether the passage in the SOP that was coupled to a guideline, is compliant with that guideline. It consists of a single Natural Language Inference (NLI) model. The output of the model is a list of probabilities for the following three NLI labels, of which the maximum probability is selected:

- **Entailment:** The SOP passage is compliant with the guideline;
- **Neutral:** The guideline is not implemented in the given SOP passage;
- **Contradiction:** The SOP passage is not compliant with the guideline.

An example of this compliance evaluation can be found in Table 1.

Table 1. Example of compliance evaluation using the three NLI labels.

Recommendation	SOP passage	NLI label
A COVID-19 patient should wear a mask.	The patient wears a mask.	Entailment
A COVID-19 patient should wear a mask.	The patient does not wear a mask.	Contradiction
A COVID-19 patient should wear a mask.	The patient wears gloves.	Neutral

Because SOPalign has been developed for healthcare in the Netherlands, we could not simply rely on a high-quality English NLI model, and had to localize our approach at this step in the pipeline by training a Dutch NLI model.

To train our Dutch NLI model, we experimented with fine-tuning three different pre-trained large language models for Dutch: RobBERT [2], BERTje [3] and mBERT [4].

We experimented with two different datasets for fine-tuning. First, we used the SICK-NL dataset [5], which is a Dutch semi-automatically translated version

of the original SICK dataset [6]. The sentences in this dataset are generated from descriptions of images. These sentences are short and not of the clinical domain, which renders these quite different from the sentences we encounter in the SOPs and guidelines. Perhaps not so surprisingly then, the model did not generalise to the sentences in our application domain.

Therefore, we translated the medNLI dataset [7] to Dutch using machine translation. This dataset was specifically created for medical texts and thus, it was a much better fit. For translation, we experimented with both Google Translate and DeepL. Table 2 shows the results on the medNLI test set for the different settings where the best model was selected after training for 20 epochs. The combination of DeepL and BERTje yields the best results for the Dutch language. The performance loss caused by the translation and the use of Dutch pre-trained models seems limited, when compared to their English counterparts.

Table 2. Results on the medNLI test set for the different translators on the three different pre-trained models.

Pre-trained model	Data language	Translator	F1-score
mBERT	English	None	79.87
mBERT	Dutch	Google	74.44
mBERT	Dutch	DeepL	75.68
BERT	English	None	79.57
BERTje	Dutch	Google	73.86
BERTje	Dutch	DeepL	76.34
RoBERTa	English	None	80.96
RobBERT	Dutch	Google	74.33
RobBERT	Dutch	DeepL	73.84

The linker and compliance modules are used as follows. We automatically generate annotations inside the SOP PDF for the guideline recommendations linked to each passage, together with the corresponding NLI label. The list of annotated PDFs is shown to the user, along with two tables, one containing all the annotations made and another with annotation counts per recommendation along with the NLI labels. The latter helps the user to find mismatches between guidelines and SOPs. The user can adapt the strictness of the system using a slider to control the threshold and influence the precision/recall trade-off. A screenshot of a hospital SOP annotated with matched guidelines is shown in Fig. 2.

The tool will be made available for hospitals and long-term care facilities to look for improvements on user experience. A built-in possibility to provide feedback on the (mis)matches identified by the tool will be used to generate new and realistic labeled data to improve performance. A limitation of our current approach is that our model is trained on machine-translated NLI data, resulting in noisy sentences. This feedback addition will overcome this. We will also calculate the Inter-Annotator Agreement (IAA) between the tool and at least two

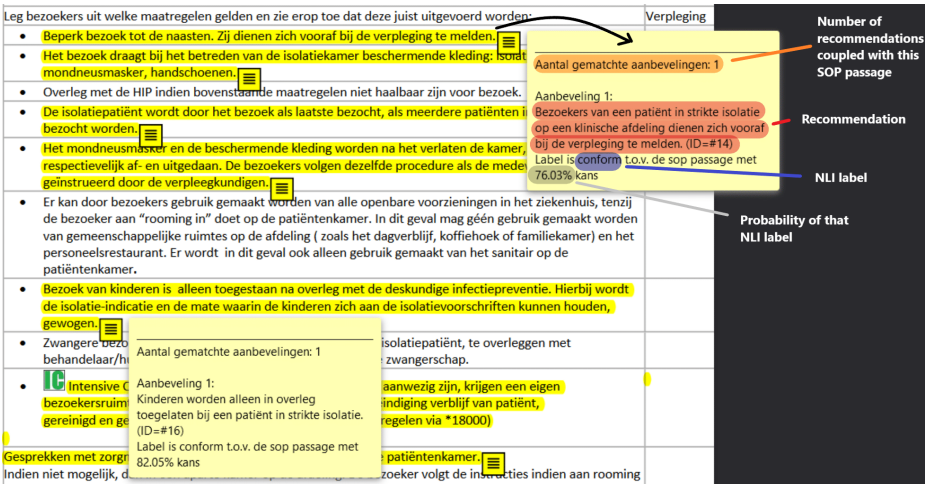

Fig. 2. Screenshot of annotated output PDF for a SOP.

human annotators to evaluate the tool, and we will perform an error analysis to look for systematic errors.

3 Demonstration

Users can access our code through GitHub[4]. The repository contains the code for the tool, the API that is used, the training of the models, the translation of the data, and the translated medNLI dataset. To showcase the important features of our tool we have created a short video[5] demonstrating an example usage of the tool where a compliance check is done of two hospital SOPs with the recommendations of the Dutch hospital MRSA guideline. The tool can easily be adapted to allow for different languages by changing the underlying STS and NLI models.

Acknowledgements This work is funded by ABR Zorgnetwerk Utrecht and Zorgnetwerk GAIN.

References

1. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In EMNLP 2019, pp. 3973—3983
2. Delobelle, P., Winters, T., Berendt, B.: RobBERT: a Dutch RoBERTa-based Language Model. In ACL 2020, pp. 3255-–3265.

[4] `https://github.com/Lukevanl/SOPalign`
[5] `https://youtu.be/vVPHGCMsmvY`

3. De Vries, W., Van Cranenburgh, A, Bisazza, A. Caselli, T. Van Noord, G. Nissim, M.: BERTje: A Dutch BERT Model. arXiv:1912.09582, 2019.
4. Devlin, J., Chang, M. -W, Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL 2019, pp. 4171—4186
5. Wijnholds, G., Moortgat, M.: SICKNL: A Dataset for Dutch Natural Language Inference. In EACL 2021, pp. 1474—1479.
6. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In LREC 2014, pp. 216-–223.
7. Romanov, A., Shivade, C.: Lessons from Natural Language Inference in the Clinical Domain. In EMNLP 2018, pp. 1586–1596.