Automatic Classification of Budget Allocation Conditions

Emma Wallerö[1], **Sven-Olof Junker**[1], **and Sara Stymne**[2]
[1]ESV, the Swedish National Financial Management Authority
[2]Department of Linguistics and Philology, Uppsala University
{emma.wallero,svenne.junker}@esv.se, sara.stymne@lingfil.uu.se

Abstract

There are indications that government's control of budget allocations has increased over the last decades. This paper attempts to examine the possibilities to automatically classify the expressed conditions of budget allocation as controlling or not. We performed an iterative annotation process in order to define the task and produce guidelines. Based on these guidelines we annotated a data set for the task. In initial experiments, we explored different classifiers and input representations. Our best system, a BERT-based classifier, reached an accuracy of 0.89, which we believe enables future large-scale studies of conditions.

1 Introduction

Language is the main means of governing and forming policies in a democracy. When conducting public administration, governments represent their decisions in terms of syntactic structures and semantics. Besides formulating laws and legislation, the controlling power normally has a wide array of textual tools to instruct public bodies in appropriate actions. In advancing the knowledge of governmental decision-making, we argue that the textuality of politics should comprise the analytical work (Hudson, 2019; Humphreys, 2021; Subramanian, 2021).

Since the 1980s, the New Public Management (NPM) paradigm has guided a large proportion of the administrative reforms in Western democracies (Peters and Pierre, 1998). Although many of the NPM reforms aimed to delegate decision autonomy to governmental agencies, studies show that NPM has also led to an increased government control of the agencies' performances (Sundström, 2003). This is referred to as the ”paradox of autonomization”, where the delegation of self-rule has been accompanied by efforts of stricter control (Wynen and Verhoest, 2016).

In 2007, a government committee in Sweden criticized the amplified control focus after almost 20 years of NPM reforms, and suggested a downplay of controlling mechanisms. The proposal, adopted in 2008, has resulted in a major reduction of the numbers of targets and reporting requirements issued to public agencies by government (ESV, 2021). Nonetheless, a recent study suggested that government in fact adopted a “balancing strategy” by adding new controlling mechanisms (Öberg and Wockelberg, 2021). Every year, the government conditions the budget allocations to all government agencies. It has been proposed that this form of budgetary control has grown since 2008, a hypothesis that still has not been tested (ESV, 2021).

This paper offers a unique step towards an advanced, big data analysis of how the textuality of politics has changed over two decades in Sweden, using language models that are trained to perform automatic classifications of over 50 000 conditions of budget allocation, which is a core part of government's budgetary control. Our aim in the paper is to examine the possibilities to automatically classify the expressed conditions of budget allocation as controlling or not.

By doing so, measurement of the possible increase or decrease in controlling conditions can be performed automatically, enabling analyses of large sets of data that would be far too extensive to classify manually. Since no previous definitions of this classification task existed, we first defined suitable categories through an iterative annotation process comprising three expert annotators. Conditions of budget allocation were classified in several rounds, followed by discussions and modifications to the guidelines. This led to consensus on which categories to use, and the construction of guidelines for the task. We then proceeded and annotated a data set for the task, with a total of

2114 examples. For the classification of conditions, we experimented with three different types of language model architectures; Support Vector Machine (SVM), a bidirectional Long Short-Term Memory Network (BiLSTM) and BERT, where we used the Swedish model KB-BERT (Malmsten et al., 2020). KB-BERT was the best performing model receiving an overall F1-score of 0.81 and and accuracy of 0.89.[1]

2 Annotation and Data

2.1 Data preprocessing

In this study, we used data consisting of expressed conditions extracted from government's annual appropriation directions between 2005 and 2022 for all public agencies. In a chapter of these directions, government conditions how the provided allocations should be used. The data was made available to us in an xlsx-file with one paragraph per line. These paragraphs were extracted from html documents, and many of the html-tags were still present in the xlsx-file. The data was split into sentences rather than any larger or smaller entity. All sentences containing less than 3 words and 15 characters were not considered since they generally, given our manual inspection, are not controlling conditions and therefore not valid for further analysis. All data inside tables was removed. If a sentence seemed to be a bullet point based on regular expressions, a certain tag ”**Listpunkt**” was concatenated to the beginning of the sentence, and extra meta data containing the final sentence of the preceding paragraph was extracted, since this was often found to be descriptive of the bullet points. The tag that indicates a bullet point was removed after the annotation process.

2.2 Annotation

To be able to perform automatic classification of conditions, we needed to annotate a data set. Since no guidelines or definitions for conditions were previously defined, we first had to define the task and create guidelines. The annotation process was performed iteratively by a master student in Language Technology (the first author) along with three employees at ESV, the Swedish Financial Management Agency (including the second author). The latter are referred to as expert annotators, since they are highly familiar with the material and task. We first performed two rounds of co-annotation, where the annotation was performed in a session with the full team who then also discussed the annotations, followed by updating the guidelines. This was followed by three additional rounds where at least two annotators annotated a sample according to the current guidelines. In a final round, the data set was annotated according to the established guidelines.

Categories
Condition controlling to the scope and content of the agency's operations **A**
Condition or information that is not controlling the agency's operations **B**
Not a condition/ cannot be labelled **C**

Table 1: Final class definitions

As for annotation tool, we used an Excel file where the annotators could see each sentence that should be tagged, along with some additional meta data. The amount of meta data was changed slightly throughout the process, but in the final round agency and context in the form of the full paragraph was presented. If the sentence was a bullet point, the last sentence in preceding paragraph was also presented.

The three classes in the final guidelines can be found in Table 1. The number of classes was changed during the annotation process. In the initial suggestions there were three to four main categories along with a number of binary subcategories. The latter considered more specified matters such as mentioning of other agency, accounting instructions, payment of funds and specification of partial sum. These subcategories were discarded after the first annotation round due to a need to focus on the main problem: detecting controlling language. In order to tell how well the annotation process was going we measured inter-annotator agreement after each individual annotation round, where the three expert annotators got the same 30 to 60 sentences to classify. By doing so, we could tell if the guidelines along with our discussion had made our agreement better or worse. The Kappa value for the first individual annotation round was 0.41. In the first round only 2 out of 3 expert annotators participated. After the second annotation round, the

[1] This work has previously been described in Wallerö (2022), where more details can be found.

	Train	Test	Valid	Total
Expert	1722	200	192	2114
Layman	339	94	38	471

Table 2: Number of annotated examples in final expert data set and layman set.

inter-annotator agreement was slightly increased, with pair-wise Kappa values ranging from 0.55 to 0.59. Some changes regarding guidelines saying that sentences containing words like ”may” and ”should” often can be labelled as ”A” or controlling were then made. These changes might not have been for the better, since the Kappa values dropped significantly for this round to be between 0.2 to 0.52. One reason for this might be that many of the randomly selected sentences were from the complex agency ”Kammarkollegiet”. It is however fortunate that this was discovered since the experts agreed on that this agency should be excluded from the study because of their complex characteristic (Wallerö, 2022).

The final Cohen’s Kappa values were 0.44, 0.55 and 0.57, for each pair of annotators. These correspond to moderate agreement (Sim and Wright, 2005). The final training set and validation set consists of 1722 and 192 sentences respectively, classified by single annotators, see Table 2. The test set consisting of 200 sentences was annotated by all three annotators individually, and the class was defined based on majority vote. The data sets are all quite skewed with around 75% of the data being labelled as A, 18% as B and 7% as C.

Apart from the expert annotated data, a set of laymen annotated data by the language technologist was used for fine-tuning of hyper-parameters. This data set is different to the final expert annotated data set in the sense that it is skewed differently, with the larger part of the data set being tagged as ”B”, not controlling. It is also created based on an earlier version of the annotation scheme than what was used in the final round by the expert annotators.

3 Experiments

In our experiments we compared how well a set of different classifiers performed on the task of classifying conditions, and we also investigated different ways of representing the input data.

3.1 Models and Experimental Setup

For our experiments we trained three different types of architectures using three different types of data input types. The architectures we used were an SVM based on the scikit-learn toolkit (Pedregosa et al., 2011), a Bi-LSTM by the open source framework Pytorch (Paszke et al., 2019) and a BERT model based on KB-BERT (Malmsten et al., 2020). For all models, grid-searches were performed for a set of hyper-parameters, based on the layman annotated data set (Wallerö, 2022). Apart from the models, a majority class baseline was presented, which always predicted the most common label which in this case was label ”A”.

The input representation used for SVM was cased including function words, which also goes for the Bi-LSTM and BERT models. Removing function words and using uncased data was tested for the SVM but yielded worse performance. The BERT model referred to is the Cased KB-BERT fine-tuned as a classifier, using the ”BertForSequenceClassification” modification.

We explored three ways of representing the input: (1) target sentence, (2), previous sentence concatenated to target sentence, and (3) agency name concatenated to target sentence. For bullet points, the last sentence of the preceding paragraph was always concatenated to the target sentence, since this context was deemed to often be crucial in order to understand the bullet points. For input representation 1 and 2, bullet points were therefore identical to each other.

The overall results of the models are presented using F1-score and accuracy, and the performance of specific classes is presented using precision and recall. The experiments presented in this article are based on the test data set. More thorough results for both validation and test sets along with a wider set of metrics can be found in the master thesis of Wallerö (2022).

4 Results

Table 3 shows overall F1-scores and accuracy for the different classifiers and input representations. The best performing model considering both F1-score and accuracy was the BERT model. The data representation type giving the best results is type 1 using no context data with the exception of previous sentence from previous paragraph for bullet points, gaining an F1-score on 0.81 and an accuracy of 0.89 using BERT. All models outperformed

	Sentence		Context+Sent		Agency+Sent	
	F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**
Baseline	0.29	0.76	0.29	0.76	0.29	0.76
SVM	0.77	0.85	0.71	0.85	0.71	0.83
BiLSTM	0.74	0.84	0.72	0.83	0.72	0.84
BERT	**0.81**	**0.89**	0.76	0.87	0.79	0.87

Table 3: Results with different input representations.

	Precision			Recall		
	A	B	C	A	B	C
Baseline	**1.0**	0.0	0.0	0.76	0.0	0.0
SVM	0.89	0.65	0.82	0.92	**0.61**	0.6
BiLSTM	0.88	0.6	0.85	0.93	0.45	0.73
BERT	**0.92**	**0.71**	**0.86**	**0.95**	**0.61**	**0.8**

Table 4: Precision and recall for each class

the Majority class baseline. The SVM model performed roughly on par with the BiLSTM model.

In Table 4, the performance of specific classes for the best performing input representation, sentence only, is presented. For all models, class A had both the highest precision and recall. This might be due to the fact that the training data set is quite skewed with around 75% of the sentences being labelled as A. 95% of all controlling sentences were correctly predicted by the BERT model, as Table 4 demonstrates a recall of 0.95. Something that is interesting to note, however, is that even though class B was more than twice as common as class C in the train and test set, the models generally gave better results for the more unusual C.

Better performances for the less represented classes could perhaps be gained by counteracting the skewedness of the data sets. This could be done by using techniques such as sampling or loss-weighting. Another solution could be to use data augmentation such as generating synthetic sentence - class pairs based on modifying existing sentences by changing words or characters and add these to the existing data sets.

In addition to the full results, we also considered the results only for bullet points, since they differ from conditions expressed as full sentences, and are always represented also with a context sentence. We found that the performance of the bullet points were generally lower than for standard sentences. This might be due to the last sentence in the preceding paragraph not being as descriptive as we thought. We think that this issue requires further exploration.

5 Conclusion

In this paper we investigated how well computational methods work for classifying controlling conditions. Since this is a new task, we started with an iterative annotation process to define the task and establish guidelines for it. Based on these guidelines we annotated a data set of 2114 conditions using three classes, which are publicly available.[2] We performed a first set of experiments on this data set, investigating different classifiers and input representations, with our best model, based on KB-BERT, reaching an overall accuracy of 0.89. The classifier developed in this study, possibly with some improvements, especially for the rarer classes, can enable large-scale studies of budget allocations which would contribute to the analysis regarding the possible changes of government controlling.

Acknowledgments

Many thanks to Annika Alexandersson and Kenneth Eliasson for your annotations and being part of developing the annotation schedule.

References

ESV. 2021. Regeringens resultatstyrning av myndigheterna - En kartläggning av instruktioner och regleringsbrev under två decennier. Report, ESV 2021:18.

Valerie M Hudson. 2019. *Artificial intelligence and international politics*. Routledge.

Ashlee Humphreys. 2021. The textuality of markets. *AMS Review*, 11(3):304–315.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden - making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Shirin Ahlbäck Öberg and Helena Wockelberg. 2021. Agency control or autonomy? government steering of

[2] https://github.com/e-wallero/classification-conditions

swedish government agencies 2003–2017. *International Public Management Journal*, 24(3):330–349.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

B Guy Peters and John Pierre. 1998. Governance without government? rethinking public administration. *Journal of public administration research and theory*, 8(2):223–243.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies use, interpretation, and sample size requirements. *Physical therapy*, 85(3).

Shivashankar Subramanian. 2021. *Natural Language Processing for Improving Transparency in Representative Democracy*. Ph.D. thesis, University of Melbourne.

Göran Sundström. 2003. *Stat på villovägar: Resultatstyrningens framväxt i ett historisk-institutionellt perspektiv*. Ph.D. thesis, Stockholm University.

Emma Wallerö. 2022. Automatic classification of conditions for grants in appropriation directions of government agencies. Master's thesis, Uppsala University.

Jan Wynen and Koen Verhoest. 2016. Internal performance-based steering in public sector organizations: Examining the effect of organizational autonomy and external result control. *Public Performance & Management Review*, 39(3):535–559.