Speech Act Classification of Swedish Sentences: Bootstrapping an embedding-based neural classifier from a rule-based classifier

Daniel Tufvesson and **Lars Ahrenberg**
Department of Computer and Information Science
Linköping University
dantu756@student.liu.se, lars.ahrenberg@liu.se

Abstract

The illocutionary force of an utterance is an important aspect of its meaning. In this work an automatic classifier for illocutionary acts in Swedish sentences was developed in three steps. First a test set was created, following the MATTER development cycle. The sentences in this test set originate from online discussion forums. Then a rule-based classifier using a subsample of the sentences was trained. Finally, this rule-based classifier was used for automatically annotating a large training set, which was then used for training a neural network for classifying speech acts—essentially bootstrapping the network from the rule-based classifier. The results show that both classifiers outperform the baseline and that the neural classifier performs better than the rule-based one. However, it was not possible to conclude if this was due to the increase in data or to its differing architecture.

1 Introduction

Modern annotated language corpora and treebanks often provide more information than a strict morpho-syntactic analysis. For example, the annotated corpora at Språkbanken give information on word senses, sentiment values and readability as provided by the Sparv pipeline (Hammarstedt et al., 2022). Information on speech acts, or even sentence types, is rarely found, however. This information is highly valuable, however, not only to enrich the corpora for theoretical language studies, but also for applications such as language teaching, and digital language assistants.

In this work we investigate the possibility of automatically annotate Swedish sentences for speech acts using a small taxonomy of general illocutionary speech acts. The most well-known taxonomy of this kind is (Searle, 1979) but here we base the classification on the overlapping taxonomy found in Svenska Akademiens Grammatik, SAG, (Teleman et al., 1999) as it has already been fitted to Swedish data.

- *Assertives* (Sw. påståenden): the speaker holds that something is true or false (or somewhere in between). For example, “Det var inte så jag menade” (’That is not what I meant’).

- *Directives* (Sw. uppmaningar): the speaker attempts to get the listener to carry out a specific action. For example: “Kan du hålla den här åt mig?” (’Will you hold this for me?’)

- Questions (Sw. frågor): the speaker requests information about whether or not something is true, or under what conditions it is true. For example: "Vad kostar bilen?" (’How much does the car cost?’)

- Expressives (Sw. värderande inställningar): the speaker expresses some feeling or emotional attitude concerning the propositional content. For example: "Vilken underbar hund!“, (’What an adorable dog!’)

In addition (Teleman et al., 1999) defines a fifth type, Hypothesis, which we did not include in this study as very few instances were found in the training data.

Specifically the following questions were investigated:

- How can an evaluation test set of sentences annotated with speech acts be created?

- How can semi-automatic methods be used for creating a large training set of sentences annotated with speech acts?

- How can a pre-trained SBERT language model be used to classify speech acts of written Swedish sentences?

2 Related work

Speech act information can be found in dedicated dialogue corpora. But the speech act classes used for dialogue corpora are often depending on a limited class of applications. This is the case with the well-known DAMSL taxonomy (Allen and Core, 1997), developed for two-agent task-oriented dialogues.

For general corpora, speech act information is rarer, but recently the Universal Dependencies project have started to investigate the annotation of abstract constructions, including speech act constructions (Weissweiler et al., 2024).

Many classifiers and classification methods have been developed, both for dialogue acts and more general speech acts. The method closest to the one used in this work is the bootstrapping method described in (Suendermann et al., 2009) where a statistical classifier for dialogue acts is bootstrapped from a rule-based classifier. The rule-based classifier consisted of handwritten rules for the grammar of the dialogue acts. It was then used for classifying utterances for the training set. The statistical classifier consisted of a trigram language model and a naive Bayes classifier. Comparing their performances, the rule-based classifier had an accuracy of 78%, while the bootstrapped statistical classifier achieved an accuracy of 90%. Hence, bootstrapping a statistical classifier from a rule-based one seems to be a viable approach.

3 Method

To obtain a speech act classifier for written sentences this work applied bootstrapping of a rule-based classifier.This was done in the following steps:

1. Swedish data was collected from corpora available at Språkbanken[1]. 15 corpora were taken from Flashback and 23 from Familjeliv. To obtain syntactic analyses in the conllu-format the data was reparsed with the Stanza parser.

2. A sample of this data with sentences from each sub-corpus was annotated manually.

3. The annotated sample was used to train a rule-based classifier employing features in the form of syntactic blocks, explained below.

4. The rule-based classifier was then used to automatically annotate a training set for the neural classifier. This training set contains more than 3,200,000 sentences.

5. Finally the two classifiers were evaluated and compared using the held-out manually annotated test data. Their performance was also compared to a baseline that classified every sentence as Assertive. Measures were computed for accuracy, recall, precision, F1, and an averaged F1.

Annotation. The annotation work largely followed the MATTER development cycle described by (Pustejovsky and Stubbs, 2012) where modelling, annotation, training, evaluation, and revision are repeated in cycles. For this study, however, only one cycle was performed. A special annotation tool was developed for the purpose. Altogether 5,450 sentences were annotated of which 2,435 were held out as a test set for the final evaluation. The classes used were Assertive, Question, Directive, Expressive, as well as two additional labels: Something_else and Unsure.

The annotation was based on guidelines that were compiled and revised iteratively. The annotation was performed by one person only but the consistency of the annotation was checked in a final round where previously annotated sentences were annotated again. To estimate the quality of the annotations, The intra-rater agreement was calculated as Cohen's kappa yielding a value of .835, which is a "Perfect" agreement according to the Landis and Koch scale of agreement, and thereby indicated that the annotation guidelines were of sufficiently good quality in terms of reliability.

The rule-based classifier. The rule-based classifier learns IF-THEN rules from the manually annotated sentences. The classifier receives a parsed sentence to be classified. It begins by identifying the root word, then picks out the words that are dependent on the root word. These words, including the root word, are then converted into syntactic blocks, (synt-blocks). The blocks each encode different kinds of grammatical information, for example, whether a word is a finite verb, or an interrogative pronoun, or a subject. Essentially, the synt-blocks encode the specific pieces of information that are relevant for determining the speech

[1] https://spraakbanken.gu.se

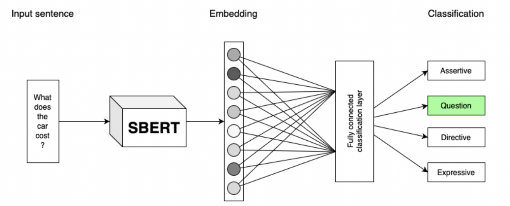

Figure 1: The neural classifier.

act of a sentence. Some words do not have a corresponding synt-block, and these are thereby not converted. For the converted words, however, their synt-blocks are placed into a sequence in the order that their corresponding words appear in the sentence. For example, the sentence “I went home late.” is converted into the synt-block sequence ’SUBJECT, FINITE_VERB, PERIOD’. This sequence thus encodes shallow syntactic information about the sentence that the system can learn from.

In addition to this, each sentence was tagged with its sentiment as positive, negative or neutral using the system developed by KBLab (Hägglöf, 2023). As the non-neutral sentiments correlated well with the Expressive class, a special post-process was added to distinguish Assertives from Expressives. A sentence classified as Assertive by the rules has its class changed to Expressive if it has a non-neutral sentiment.

Rules are ordered based on different parameters. When new sentences are annotated their synt-block sequence is matched against the rules in their order. The match need not be perfect but every synt-block of the sentence need to be found in the rule.

The neural classifier. This classifier consists of a sentence embedding layer and a linear classification layer (see Figure 1). A sentence is first fed to the embedding layer where it is computed into a sentence embedding, a Swedish SBERT transformer model, pre-trained by KBLab (Rekathati, 2021). SBERT was chosen because it produces semantically meaningful sentence embeddings. The embedding is next fed to the classification layer, a single, fully connected, linear layer, that classifies the embedding with the speech act that receives the highest score.

The training was done solely on the classification layer by back-propagating the loss gradients. The loss was computed with a categorical cross-entropy loss function. Since the training data was unbalanced, the loss calculation was done in conjunction with class weights to counteract the effects of the

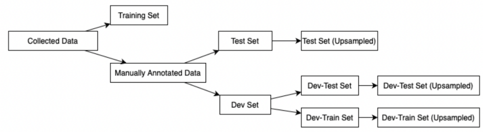

Figure 2: The corpus with splits for different uses.

unbalanced classes. The training was done over 10 epochs, with a learning rate of .01, and a batch size of 16 using the Adam optimizer.

The classifier models and code are available on GitHub[2] and the data sets are available on Kaggle[3].

4 Results

In total, eight data sets were created, distributed on train- development- and test sets for the different steps in the process. These are shown in 2.

As illustrated in Table 1, the embedding-based classifier achieves both the highest accuracy and average F1-score. Both the rule-based and embedding-based classifiers achieve higher scores than the baseline. The embedding-based classifier has a .05 higher accuracy and a 0.4 higher F1- score than the rule-based one.

	Baseline	**Rule-based**	**Neural**
Accuracy	0.25	0.69	**0.74**
Avg. F1	0.10	0.70	**0.74**

Table 1: Accuracy and average F1-scores for the two classifiers and the baseline. Best scores in bold.

Table 2 shows performance on individual classes. We can see there that the rule-based classifier sometimes has a better score than the neural one, in 4 out of 12 metrics. We also see that Questions are by far the class giving the best performance, while Assertives and Directives are generally the hardest to recognize for the systems.

5 Discussion

As for developing a test set, the results show that it can be done with the MATTER development cycle. Some deviations were necessary for the limited scope of this work; only one person annotated the sentences and only one cycle of annotation was performed. As described above, intra-annotator agreement was checked in a final round of annotation and was found to be high.

[2]https://github.com/Daniel-B-Tufvesson/speech-act-classifier

[3]https://www.kaggle.com/datasets/danieltufvesson/swedics-speech-acts

	Precision		Recall		F1-score	
	(Rule-based)	(Neural)	(Rule-based)	(Neural)	(Rule-based)	(Neural)
Assertives	0.53	**0.60**	**0.74**	0.70	0.62	**0.64**
Questions	**0.96**	**0.94**	0.92	**0.93**	0.94	0.92
Directives	**0.76**	0.72	0.60	**0.75**	0.67	**0.73**
Expressives	0.64	**0.72**	0.51	**0.57**	0.57	**0.63**

Table 2: Class-specific performance of the two classifiers. Best scores in bold face.

The rule-based classifier employing syntactic blocks and sentiment values worked very well. It vastly outperformed the baseline on both accuracy and F1-score. It can differentiate between speech acts, albeit with some exceptions for directives and expressives. The most likely reason is that they are expressed indirectly. The rule-based classifier was then used for automatically annotating a large training set of sentences.

The classifier requires sentences to be annotated with part-of-speech tags, lemmas, dependency relations, morphology, and sentiment—all of which can be done automatically. Hence, the process of creating a training set can in principle be fully automated—from data pre-processing to speech act annotation—with the only exception of actually training the rule-based classifier.

The Swedish SBERT model computed the embeddings of the sentences, which were then classified with a single, fully connected, linear layer. This classifier achieved a higher accuracy and F1 than the rule-based classifier with the greatest improvements for expressives and directives. This suggests that the neural classifier has learned the semantic differences between speech acts enough to better differentiate between them. It also shows that the semi-supervised learning approach taken in this work is a viable method for creating an neural classifier, that is, first training a rule-based classifier on a small set of manually annotated sentences, and then using the classifier to automatically annotate a larger training set.

We have not checked what performance the neural classifier would achieve if it was instead trained only on the dev-train set. Thus, it is not possible to conclude that the bootstrapping is improving the final performance. It can only be concluded that the knowledge in the rule-based classifier can be distilled into an embedding-based neural network classifier—going from syntax to semantics.

References

James Allen and Mark Core. 1997. Draft of damsl: Dialogue act markup in several layers.

Martin Hammarstedt, Anne Schumacher, Lars Borin, and Markus Forsberg. 2022. *Sparv 5 User Manual*. Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet.

Hillevi Hägglöf. 2023. The kblab blog: A robust, multi-label sentiment classifier for swedish.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

John R. Searle. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

David Suendermann, Keelan Evanini, Jackson Liscombe, Phillip Hunter, Krishna Dayanidhi, and Roberto Pieraccini. 2009. From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 4713–4716. IEEE.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens Grammatik, Vol. 4*. Svenska Akademien.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.