Bridging the gap between sign language machine translation and sign language animation using sequence classification

Sarah Ebling[1], *Matt Huenerfauth*[2]

[1]Institute of Computational Linguistics
University of Zurich
Zurich, Switzerland
ebling@cl.uzh.ch
[2]Rochester Institute of Technology (RIT)
Golisano College of Computing and Information Sciences
Rochester, NY, USA
matt.huenerfauth@rit.edu

Abstract

To date, the non-manual components of signed utterances have rarely been considered in automatic sign language translation. However, these components are capable of carrying important linguistic information. This paper presents work that bridges the gap between the output of a sign language translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. More precisely, the generation of non-manual information is scheduled after the machine translation step and treated as a sequence classification task. While sequence classification has been used to solve automatic spoken language processing tasks, we believe this to be the first work to apply it to the generation of non-manual information in sign languages. All of our experimental approaches outperformed lower baseline approaches, consisting of unigram or bigram models of non-manual features.

1. Introduction

Sign languages are often the preferred means of communication of deaf and hard-of-hearing persons, making it vital to provide access to information in these languages. Technologies for automatically translating written text (in a spoken language[1]) into a sign language would therefore increase the accessibility of information sources for many people.

Sign languages are natural languages and, as such, fully developed linguistic systems. While there are a variety of sign languages used internationally, they share several key properties: Utterances in sign languages are produced with the hands/arms (the *manual activity*) and the shoulders, head, and face (the *non-manual activity*). Manual and non-manual components together form the *sublexical components*.

1.1. Sign language production pipeline

While the input to a translation system such as the one outlined above would be a written text, the output is less obvious: Ultimately, the goal would be to produce an animation of a virtual human character performing sign language, i.e., a *sign language avatar*. Most sign language machine translation systems produce some form of symbolic output. In the ideal case, this output should be suitable to serve as the *input* for an animation-synthesis system.

Unfortunately, to date, this sign language production pipeline is often left incomplete, in that the output of many machine translation systems consists of strings of sign language glosses,[2] i.e., information about the manual activity of a signed utterance, only.

This paper presents work that *bridges the gap* between the output of a sign language translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. More precisely, the generation of non-manual information is scheduled after the machine translation step and treated as a sequence classification task. To our knowledge, this is the first work to apply sequence classification to the generation of non-manual information in sign languages. We show that all of our experimental approaches outperformed lower baseline approaches, including unigram and bigram models of non-manual component sequences.

1.2. Linguistic background and prior work

Experimental research with sign language users has shown that the absence of non-manual information in synthesized signing (sign language animation) leads to lower comprehension scores and lower subjective ratings of the animations [1]. This is because non-manual components in sign languages are capable of assuming functions at all linguistic levels [2]. As an example, in Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS), raised eyebrows are used to express supposition, contrast, or emphasis [3]. A combination of a head movement forward and raised eyebrows is used to mark topicalized constituents. Conditional *if/when* utterances have the head tilt and move forward slightly and the eyebrows go up at the start of the condition part. For rhetorical questions in DSGS, the head tilts and moves forward slightly and the eyebrows are furrowed on the question sign [4].

Non-manual components have been omitted, for example, in a statistical machine translation system that translates between German and German Sign Language [5] and one that

[1] The term *spoken language* refers to a language that is not signed, whether it is represented in its spoken or written form.

[2] Glosses are semantic representations of signs that typically take on the base form of a word in the surrounding spoken language.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 2–9,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

translates between English, German, Irish Sign Language, and German Sign Language [6]. Massó and Badia [7] took into account mouth morphemes in a statistical machine translation system translating from Catalan into Catalan Sign Language; such mouth movements convey adverbial or aspectual modifications to the meaning of manual signs in that language.

In contrast, in this paper, we deal with multiple types of non-manual components, taking into account the multilinear nature of sign languages. Our work is inspired by linguistic models that represent both the manual and non-manual components of signed utterances [8, 9].

The remainder of the paper is structured as follows: Section 2 introduces the project as part of which the machine translation system is being developed. In particular, the data that served as a basis for the sequence classification experiments is described. In Section 3, we specify our sequence classification approaches, provide further information on the data used for the experiments, explain the experiment configurations, and present as well as discuss the results.

2. Non-manual components in a corpus of DSGS train announcements

We are developing a system that automatically translates written German train announcements of the Swiss Federal Railways into DSGS. Our team includes Deaf[3] and hearing researchers. Example 1 below shows an announcement of the Swiss Federal Railways.

(1) *Ausfallmeldung zur S1 nach Luzern: Die S1 nach Luzern, Abfahrt um 6 Uhr 10, fällt aus.* ('Notice of cancellation regarding the S1 to Lucerne: The S1 to Lucerne, scheduled to leave at 6:10am, has been cancelled.')

The resulting DSGS announcements are presented by means of an avatar. A state-of-the art avatar system, JASigning [11], is used for this. The JASigning character *Anna* is shown in Figure 1.

The train announcements of the Swiss Federal Railways are parametrized in that they are based on templates with slots, where slots are, e.g., the names of train stations, types of trains, or reasons for delays. When automatically translating these announcements, one possibility is to take account precisely of their parametrized nature. However, our goal is to build a translation system that can later be extended to other domains with more lexical and syntactic variation. Hence, a more transferrable translation approach is applied, namely statistical machine translation.

Statistical machine translation systems require parallel corpora as their training, development, and test data. To build a parallel corpus, the Deaf and hearing members of our team manually translated 3000 written German train announcements into DSGS. The DSGS side of the resulting parallel corpus consists of information arranged on three tiers:

1. sign language glosses
2. head, with 13 possible values
3. eyebrows, with 3 possible values

The non-manual components in the DSGS side of our parallel corpus serve various linguistic functions. For example, in our domain of train announcements, we have observed that furrowed eyebrows often occurred during signs with negative polarity, such as the sign BESCHRÄNKEN ('LIMIT'). Raised eyebrows often occurred during signs that express a warning or emphasis, e.g., the signs VORSICHT ('CAUTION') or SOFORT ('IMMEDIATELY'). The syntactic functions mentioned in Section 1.2, topicalization and rhetorical question, also occur frequently in the corpus; a few instances of conditional expressions are also present. Many of these syntactic non-manuals relate to specific words in the sentence (e.g., rhetorical question non-manual components co-occur with question words, such as "WHAT"). Within this paper, we focus on such *lexically-cued non-manuals*. (As discussed in Section 4, we are aware that not all non-manual components are predictable based on the sequence of lexical items in the sentence alone, and we propose to investigate such non-manuals in future work.)

Table 1 shows the DSGS translation of the first part of the train announcement introduced in Example 1, *Ausfallmeldung zur S1 nach Luzern* ('Notice of cancellation regarding the S1 to Lucerne'). Note that the starting and ending times of the non-manual components align with the boundaries of manual activities (as represented through glosses). This has been shown to be the case for non-manual components with linguistic functions; non-manual components that serve purely affective purposes, e.g., expressing anger or disgust, are known to start slightly earlier than the surrounding manual components [12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

3. Generating non-manual information through sequence classification

The goal of our work was to include non-manual information in the process of translating written German train announcements into DSGS. Traditionally, glosses have been the sole representation of sign language in an automatic translation task (cf. Section 1). One way of considering non-manual components in this task is to simply append them to the glosses. This representation is shown in Example 2 for the announcement introduced in Example 1. The non-manual features are printed in bold.

(2) *Ausfallmeldung zur S1 nach Luzern*:
MELDUNG**__Head_forward__Eyebrows_raised**
IX**__Head_back__Eyebrows_raised**
BAHN**__Head_up__Eyebrows_raised**
S1**__Head_down__Eyebrows_raised**
NACH**__Head_up__Eyebrows_neutral**
LUZERN**__Head_up__Eyebrows_raised**
AUSFALL**__Head_down__Eyebrows_raised**

However, such a representation aggravates the issue of data sparseness, since the size of the vocabulary is not only equivalent to the number of unique glosses but to the number of unique combinations of glosses and non-manual features. This increases the likelihood that tokens appear in the decoding phase that have not been seen during training (*out-of-vocabulary items*, OOV). Such a representation also does not accommodate the multimodal nature of sign languages: Three tiers (glosses, head, and eyebrow information) are collapsed into one.

We propose an approach that schedules the automatic generation of non-manual information after the machine translation step and views it as a sequence classification task. This is justified by the fact that the non-manual components in our corpus

[3] It is a widely recognized convention to use the upper-cased word *Deaf* for describing members of the linguistic community of sign language users and, in contrast, to use the lower-cased word *deaf* when describing the audiological state of a hearing loss [10].

Glosses	MELDUNG ('NOTICE')	IX ('IX')	BAHN ('TRAIN')	S1 ('S1')	NACH ('TO')	LUZERN ('LUCERNE')	AUSFALL ('CANCELLATION')
Eyebrows	raised				neutral	raised	
Head	forward	back	up	down	up		down

Table 1: DSGS translation of *Ausfallmeldung zur S1 nach Luzern* ('Notice of cancellation regarding the S1 to Lucerne')

serve linguistic functions, which means their boundaries align with those of manual components (cf. Section 2). Hence, the process of generating non-manual components can be regarded as a task of labeling glosses (as representations of the manual components) with non-manual features.

Figure 1 visualizes the overall pipeline that transforms a written German train announcement into a DSGS animation: The machine translation system receives as input a German announcement like the one introduced in Example 1. With the help of models learned from our parallel corpus, the system translates the German announcement into DSGS glosses. The glosses in turn serve as input for the sequence classification system. The output of the machine translation and the sequence classification system is then combined and converted into motion data for the avatar. The process of generating the motion data is not illustrated further in the figure, as it is outside of the scope of this paper.

3.1. Conditional Random Fields

Sequence classification has been used to solve various natural language processing problems, such as part-of-speech tagging and chunking (shallow parsing). In contrast to standard classifiers, sequence classifiers are capable of taking into account the sequential nature of data. Sequential Conditional Random Fields (CRFs) [22] are a state-of-the-art approach for this. Given one or more sequences of tokens (the *evidence*), CRFs compute the probability of a sequence of labels (the *outcome*). While multiple evidence layers are permitted, CRFs only allow the prediction of one outcome layer.

The Wapiti toolkit [23] provides an efficient implementation of CRFs.[4] Sequence classification with Wapiti follows a train–test–evaluate cycle. Handcrafted *feature templates* are created to specify which tokens of the evidence are considered for the prediction of the outcome labels. In addition, the *emission order* is declared, indicating whether the evidence is conditioned on label unigrams (emission order 1) or bigrams (emission order 2). During the training step, the feature templates are instantiated with the training data. Wapiti offers a model dump function, which allows the user to investigate the quality of the resulting features.

3.2. Data

To perform the sequence classification experiments in Wapiti, the parallel corpus of 3000 German/DSGS train announcements described in Section 2 was randomly divided into ten folds of 300 sentences each to enable ten-fold cross validation. For each validation round, eight folds were used for training, one was used for development, and one for testing. Using the ground truth as opposed to the machine translation output (cf. Section 2) as data was motivated by our interest in investigating the potential of sequence classification in isolation, without possible error propagation from the preceding machine translation step.

3.3. Experiment configurations

The goal of the experiments described here was to predict the most probable sequence of non-manual features for a sequence of glosses output by the machine translation system (cf. Figure 1). As stated in Section 3.1, CRFs allow the prediction of one outcome layer at a time. Hence, the two label layers head and eyebrows in our corpus (cf. Section 2) could either be collapsed into a single label (Configuration G→H+E, Table 2), or a separate classifier could be trained for each feature (Configurations G→H and G→E, Table 3). A downside of Configuration G→H+E is that there is a potential for data sparseness, as the number of possible outcome labels is equivalent to the number of cross-combinations of head and eyebrow labels occurring in the training data. However, even with this approach, the risk of data sparseness is lower than that of appending the non-manual features to the sign language glosses during the machine translation task, as described at the beginning of Section 3.

Evidence **Gloss**	**Label** **Non-manual**
MELDUNG ('NOTICE')	forward_raised
IX ('IX')	back_raised
BAHN ('TRAIN')	up_raised
S1 ('S1')	down_raised
NACH ('TO')	up_neutral
LUZERN ('LUCERNE')	up_raised
AUSFALL ('CANCELLATION')	down_raised

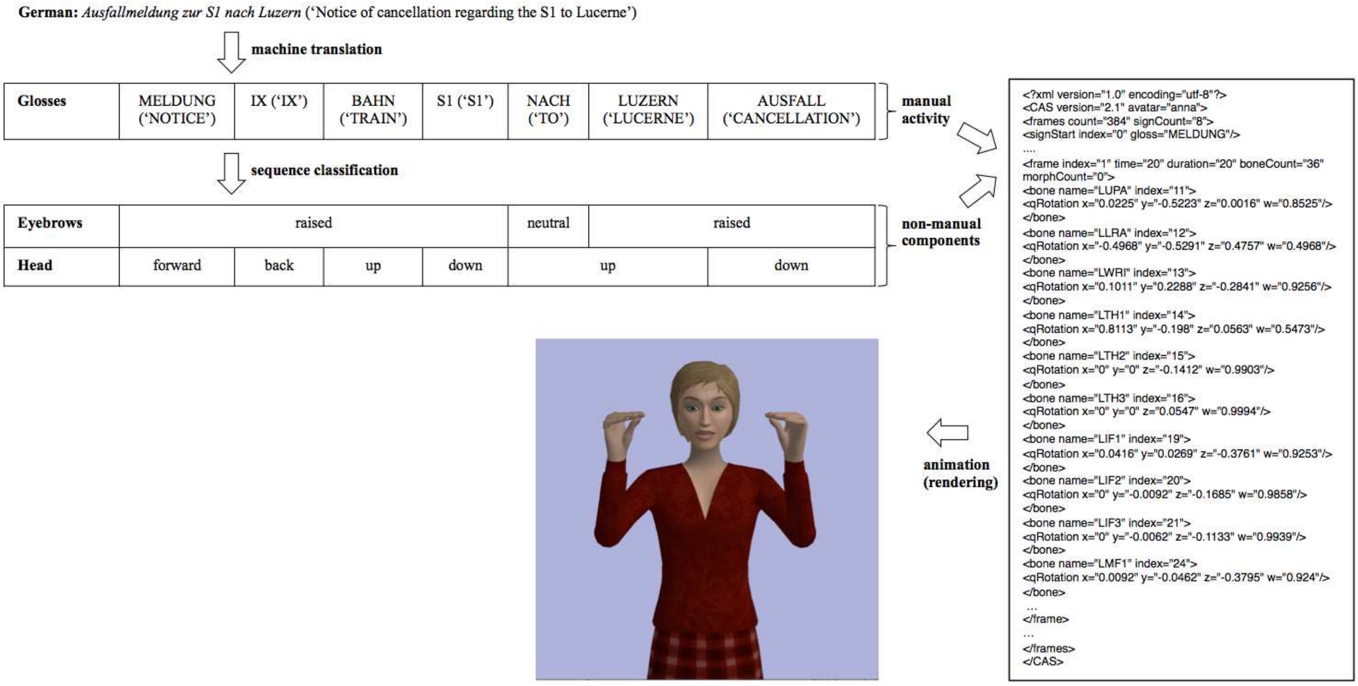
Table 2: Configuration G→H+E

Evidence **Gloss**	**Label** **Head**
MELDUNG ('NOTICE')	forward
IX ('IX')	back
BAHN ('TRAIN')	up
S1 ('S1')	down
NACH ('TO')	up
LUZERN ('LUCERNE')	up
AUSFALL ('CANCELLATION')	down

Evidence **Gloss**	**Label** **Eyebrow**
MELDUNG ('NOTICE')	raised
IX ('IX')	raised
BAHN ('TRAIN')	raised
S1 ('S1')	raised
NACH ('TO')	neutral
LUZERN ('LUCERNE')	raised
AUSFALL ('CANCELLATION')	raised

Table 3: Configurations G→H (top) and G→E (bottom)

[4] `http://wapiti.limsi.fr/manual.html`



Glosses	MELDUNG ('NOTICE')	IX ('IX')	BAHN ('TRAIN')	S1 ('S1')	NACH ('TO')	LUZERN ('LUCERNE')	AUSFALL ('CANCELLATION')
Eyebrows	raised				neutral	raised	
Head	forward	back	up	down	up		down

```
<?xml version="1.0" encoding="utf-8"?>
<CAS version="2.1" avatar="anna">
<frames count="384" signCount="8">
<signStart index="0" gloss="MELDUNG"/>
....
<frame index="1" time="20" duration="20" boneCount="36"
morphCount="0">
<bone name="LUPA" index="11">
<qRotation x="0.0225" y="-0.5223" z="0.0016" w="0.8525"/>
</bone>
<bone name="LLRA" index="12">
<qRotation x="-0.4968" y="-0.5291" z="0.4757" w="0.4968"/>
</bone>
<bone name="LWRI" index="13">
<qRotation x="0.1011" y="0.2288" z="-0.2841" w="0.9256"/>
</bone>
<bone name="LTH1" index="14">
<qRotation x="0.8113" y="-0.198" z="0.0563" w="0.5473"/>
</bone>
<bone name="LTH2" index="15">
<qRotation x="0" y="0" z="-0.1412" w="0.9903"/>
</bone>
<bone name="LTH3" index="16">
<qRotation x="0" y="0" z="0.0547" w="0.9994"/>
</bone>
<bone name="LIF1" index="19">
<qRotation x="0.0416" y="0.0269" z="-0.3761" w="0.9253"/>
</bone>
<bone name="LIF2" index="20">
<qRotation x="0" y="-0.0092" z="-0.1685" w="0.9858"/>
</bone>
<bone name="LIF3" index="21">
<qRotation x="0" y="-0.0062" z="-0.1133" w="0.9939"/>
</bone>
<bone name="LMF1" index="24">
<qRotation x="0.0092" y="-0.0462" z="-0.3795" w="0.924"/>
</bone>
 ...
</frame>
...
</frames>
</CAS>
```

Figure 1: Sign language production pipeline: machine translation, sequence classification, and animation

With Configurations G→H and G→E, each label layer (head and eyebrows, respectively) is treated in isolation, which means that dependencies between the two are not captured. However, conceptually, dependencies between these two types of non-manual features exist in that they assume specific linguistic functions together, e.g., topicalization, rhetorical questions, or conditional expressions in DSGS. These dependencies can be accounted for by introducing a cascaded approach, i.e., by using the output of one classifier as additional input for the other. More precisely, the output of the head information classifier can be used as additional evidence for the eyebrow information classifier and vice versa. This is shown as Configurations G_E→H and G_H→E in Table 4. Note that such a representation accommodates the multi-tier nature of sign languages.

To better model the sequential dependencies in a given data set, an IOB representation [24] could be used. In this format, B denotes the first token of a label sequence, I a sequence-internal token, and O is used for tokens that are not part of a sequence of a label under consideration. This format is applied as Configurations G→H$_{\text{IOB}}$ and G→E$_{\text{IOB}}$ (Table 5). Note that in the case at hand, O does not occur, since the data contains multi-class as opposed to binary annotations and *neutral* is one of the possible class labels.

For our experiments, we applied all of the above seven configurations, as summarized in Table 6.

Among the strengths of CRFs is their ability to handle a large amount of features as well as to cope with redundancy [23]. We provided 26 feature templates for each evidence layer, guided by a template provided by Roth and Clematide [25]. Table 7 lists the 12 context windows used. The table shows that the overall context considered ranged from the three previous tokens to the three following tokens relative to the current position. Each window was included with both emission order 1 (unigram) and 2 (bigram). In addition, raw unigram output distribution and bigram output distribution were included.

Evidence		**Label**
Gloss	**Eyebrows**	**Head**
MELDUNG ('NOTICE')	raised	forward
IX ('IX')	raised	back
BAHN ('TRAIN')	raised	up
S1 ('S1')	raised	down
NACH ('TO')	neutral	up
LUZERN ('LUCERNE')	raised	up
AUSFALL ('CANCELLATION')	raised	down

Evidence		**Label**
Gloss	**Head**	**Eyebrows**
MELDUNG ('NOTICE')	forward	raised
IX ('IX')	back	raised
BAHN ('TRAIN')	up	raised
S1 ('S1')	down	raised
NACH ('TO')	up	neutral
LUZERN ('LUCERNE')	up	raised
AUSFALL ('CANCELLATION')	down	raised

Table 4: Configurations G_E→H (top) and G_H→E (bottom)

3.4. Results

Table 8 shows the results of our experiments obtained using the default settings of Wapiti. "Experimental approach" refers to the configurations described in Section 3.3. The lower baseline for each configuration consisted of using the unigram ("Lower baseline, unigram") and bigram ("Lower baseline, bigram") output distribution of the labels, respectively, i.e., of globally assigning the most frequent label unigram and bigram of the training data, respectively. For each experimental or baseline

Evidence	**Label**
Gloss	**Head**
MELDUNG (‘NOTICE’)	B_forward
IX (‘IX’)	B_back
BAHN (‘TRAIN’)	B_up
S1 (‘S1’)	B_down
NACH (‘TO’)	B_up
LUZERN (‘LUCERNE’)	I_up
AUSFALL (‘CANCELLATION’)	B_down

Evidence	**Label**
Gloss	**Eyebrow**
MELDUNG (‘NOTICE’)	B_raised
IX (‘IX’)	I_raised
BAHN (‘TRAIN’)	I_raised
S1 (‘S1’)	I_raised
NACH (‘TO’)	B_neutral
LUZERN (‘LUCERNE’)	B_raised
AUSFALL (‘CANCELLATION’)	I_raised

Table 5: Configurations G→H$_{IOB}$ (top) and G→E$_{IOB}$ (bottom)

Configuration	**Evidence**	**Label**
G→H+E	glosses	head and eyebrows
G→H	glosses	head
G→E	glosses	eyebrows
G_E→H	– glosses – eyebrows	head
G_H→E	– glosses – head	eyebrows
G→H$_{IOB}$	glosses	head `IOB`
G→E$_{IOB}$	glosses	eyebrows `IOB`

Table 6: Overview of configurations

Relative position	**Description**
0	current token
-1 to +1	previous, current, following token
-1 to 0	previous and current token
0 to +1	current and following token
-1	previous token
1	following token
-2 to 0	two previous and current token
-2 to +1	two previous, current, following token
-3 to 0	three previous and current token
-2 to +2	two previous, current, following token
-1 to +2	previous, current, two following tokens
-1 to +3	previous, current, three following tokens

Table 7: Context windows used for the feature templates

approach, Table 8 provides the following numerical information, in analogy to previous work in classification for natural language processing [26, 25, 27]:

- Number of labels
- Token error: This is the mean of the token errors of the ten rounds of a 10-fold cross validation. The token error for an individual validation round is calculated as the percentage of incorrectly predicted labels.
- Standard deviation of token error for the ten rounds
- Confidence interval of token error: This is the confidence interval at a confidence level of 95% calculated over the mean of the token errors using Student’s t-test.
- Sequence error: This is the mean of the sequence errors of a 10-fold cross validation. The sequence error for an individual validation round is calculated as the percentage of incorrectly predicted sequences, i.e., sequences containing at least one token error.
- Standard deviation of sequence error
- Confidence interval of sequence error: This is the confidence interval at a confidence level of 95% calculated over the mean of the sequence errors using Student’s t-test.

The results in Table 8 show that all experimental approaches outperformed their lower baselines (unigram and bigram output distribution); in all cases, the magnitude of the difference was greater than the confidence interval of the values. The error rates of the experimental approaches are notably low, which is at least partly due to the nature of the data used for the experiments: As described in Section 2, the train announcements are highly parametrized in that they are based on a limited set of phrasal templates. As shown in the table, there is a tendency for the bigram baseline to perform better than the unigram baseline, a result that underlines the inherently sequential nature of the data.

3.4.1. Cascaded vs. non-cascaded

Between Configuration G→H (non-cascaded) and G_E→H (cascaded), both predicting head information, Configuration G→H exhibits a lower sequence error rate. Between Configuration G→E (non-cascaded) and G_H→E (cascaded), both predicting eyebrow information, Configuration G_H→E achieved a lower sequence error rate.

To examine the theoretical potential of the cascaded approach, we determined the upper bound, i.e., the result of applying the model learned from the training data on the ground-truth data. In other words, as data for the additional evidence layer (eyebrow information for Configuration G_E→H and head information for Configuration G_H→E), we used the gold-standard annotations of these layers instead of the output of Configurations G→H and G→E, respectively. The resulting numbers are shown in the table as “Upper bound” for Configurations G_E→H and G_H→E: Configuration G_E→H/Upper bound achieved a lower sequence error rate than Configuration G→H. Configuration G_H→E/Upper bound also achieved a lower sequence error rate than Configuration G→E; here, the magnitude of the difference is greater than the confidence intervals of the values. These results show that a cascaded approach is capable of outperforming a non-cascaded approach, and they imply that in DSGS, head information provides more useful information for predicting eyebrow information than vice versa.

Configuration	Labels	Token level			Sequence level		
		Token error (%)	Standard deviation	Confidence interval	Sequence error (%)	Standard deviation	Confidence interval
Predicting H+E	**31**						
Lower baseline, bigram		65.04	0.89	0.63	99.97	0.10	0.07
Lower baseline, unigram		68.16	0.51	0.36	100.00	0.00	n.a.
G→H+E		1.88	0.50	0.36	10.43	2.53	1.81
Predicting H	**13**						
Lower baseline, bigram		63.65	1.05	0.75	99.97	0.10	0.07
Lower baseline, unigram		68.07	0.50	0.36	100.00	0.00	n.a.
G→H		1.62	0.45	0.32	8.96	2.43	1.7
G_E→H		1.62	0.50	0.36	9.19	2.40	1.72
— Upper bound		1.29	0.39	0.28	7.86	2.05	1.46
Predicting E	**3**						
Lower baseline, bigram		20.57	1.07	0.77	94.62	1.80	1.29
Lower baseline, unigram		20.57	1.07	0.77	94.62	1.80	1.29
G→E		0.74	0.24	0.17	6.85	1.81	1.29
G_H→E		0.66	0.16	0.12	5.72	0.96	0.69
— Upper bound		0.45	0.11	0.08	4.21	0.88	0.63
Predicting $\mathbf{H}_{\mathbf{IOB}}$	**21**						
Lower baseline, bigram		74.75	3.98	2.85	99.97	0.10	0.07
Lower baseline, unigram		76.41	0.52	0.37	100.00	0.00	n.a.
G→$\mathrm{H}_{\mathrm{IOB}}$		1.81	0.56	0.40	9.13	2.84	2.03
Predicting $\mathbf{E}_{\mathbf{IOB}}$	**6**						
Lower baseline, bigram		37.07	1.55	1.11	95.79	1.58	1.13
Lower baseline, unigram		43.12	1.48	1.06	98.83	0.48	0.34
G→$\mathrm{E}_{\mathrm{IOB}}$		1.41	0.30	0.21	9.96	1.85	1.33

Table 8: Sequence classification experiments: Results

3.4.2. `IOB` *vs. non-*`IOB`

Between Configuration G→H (non-`IOB` format) and G→$\mathrm{H}_{\mathrm{IOB}}$ (`IOB` format), both predicting head information, Configuration G→H produced a lower sequence error rate. Between Configuration G→E (non-`IOB` format) and G→$\mathrm{E}_{\mathrm{IOB}}$ (`IOB` format), both predicting eyebrow information, Configuration G→E yielded a lower sequence error rate. In this case, the magnitude of the difference was greater than the confidence interval of the values. These results show that applying an `IOB` format was not beneficial for the task at hand, most likely due to data sparseness: Introducing the `IOB` format doubled the number of labels for Configuration G→$\mathrm{E}_{\mathrm{IOB}}$ compared to Configuration G→E (6 vs. 3 labels), while the relative increase was lower for Configuration G→$\mathrm{H}_{\mathrm{IOB}}$ compared to Configuration G→H (21 vs. 13 labels), indicating that five head information features appeared sequence-initially only, i.e., spanned over one gloss.

3.4.3. *Analysis of features*

We examined the 50 highest-weighted (instantiated) features in the models of the experimental approaches of Configurations G→H+E, G_E→H, and G_H→E for the first round of the 10-fold cross validation: Among the highest-weighted features for Configuration G→H+E were 31 bigram features and 19 unigram features. The most frequently occurring feature context window (cf. Table 7) consisted of the current token of the (gloss) evidence layer (i.e., relative position 0). Thus, the identity of a lexical item contributed to the model's prediction of the non-manual feature that co-occurs with it. The second- and third-best performing feature context windows contained the previous token (-1) and the following token (+1) of an evidence layer, respectively, and this was followed by a window containing the current and the following token (0 to +1). Thus, the neighboring lexical items contributed to the prediction of the non-manual feature.

For Configuration G_E→H (predicting head information), the 50 top-weighted features consisted of 26 bigram and 24 unigram features. 48 features used tokens from the gloss evidence layer, while 2 used tokens from the added eyebrow information layer. For Configuration G_H→E (predicting eyebrow information), this number was considerably higher: Among the 50 best-scoring features were 27 that used tokens from the head information layer. Again, this serves as evidence that head information is valuable when predicting eyebrow information in DSGS.

4. Conclusion and outlook

We have presented work that bridges the gap between the output of a sign language machine translation system and the input of a sign language animation system by incorporating non-manual information into the final output of the translation system. This is in contrast to many prior statistical sign language machine translation approaches that did not include non-manual components in their output. The inclusion of such non-manual information enables the final animation-synthesis component of a translation system to control the head and face of a signing avatar.

Our approach has scheduled the generation of non-manual information after the machine translation task and treated it as a sequence classification task. This is justified by the fact that the boundaries of linguistic non-manual components align with those of manual components, rendering the process of generating non-manual components a gloss-labeling task.

Sequence classification is a technique commonly used in the automatic processing of spoken languages. The work we have reported on in this paper is presumably the first to apply it to the generation of non-manual information in sign languages.

We have shown that all of our experimental approaches outperformed their lower baselines. The experimental approaches consisted of: predicting head and eyebrow information together in one label, predicting head and eyebrow information separately, predicting head information by using eyebrow information as additional evidence and vice versa (cascaded approach), and applying an `IOB` format. We have demonstrated the potential of applying a cascaded approach, and we did not observe any benefit from utilizing an `IOB` format in our task.

As a next step, we intend to apply our experimental approaches to a parallel English/American Sign Language (ASL) corpus from a domain with greater semantic and syntactic variety [28]. In this way, we hope to determine, for a different sign language and for a more diverse corpus, if the key findings of this paper are replicable, namely: (a) that the non-manual components of a sign language sentence may be successfully predicted using a sequence classification approach and (b) that some orderings of cascading the sequential predictions are more successful than others (e.g., for DSGS, we found that head information was useful to consider when predicting eyebrow information).

In contrast to the train announcement DSGS corpus used in this paper, the ASL corpus is known to contain instances of non-manual information that are not lexically-cued, i.e., they are not recoverable from the glosses alone. For example, a *yes/no* question in ASL can have the same surface form (gloss order) as a declarative sentence.[5] Thus, we anticipate investigating how to best exploit information contained in the (English) source sentence, e.g., to include question marks as absolute features. Leveraging information from the source sentence will also allow us to capture instances in which a syntactic function is expressed non-manually only on the ASL side: e.g., in ASL, it is possible to convey negation solely via headshake, without the use of any manual sign to indicate the negation [29].

5. References

[1] H. Kacorri, P. Lu, and M. Huenerfauth, "Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation," *ACM Transactions on Accessible Computing*, vol. 5, 2013.

[2] O. Crasborn, "Nonmanual Structures in Sign Language," in *Encyclopedia of Language & Linguistics*, 2nd ed., K. Brown, Ed. Oxford: Elsevier, 2006, vol. 8, pp. 668–672.

[3] C. Steiner, *Über die Funktion des Anhebens der Augenbrauen in der Deutschschweizerischen Gebärdensprache DSGS*, ser. Informationsheft. Zürich: Verein zur Unterstützung des Forschungszentrums für Gebärdensprache, 2000, vol. 35.

[4] P. Boyes Braem, *Einführung in die Gebärdensprache und ihre Erforschung*, 3rd ed., ser. Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser. Hamburg: Signum, 1995, vol. 11.

[5] D. Stein, C. Schmidt, and H. Ney, "Analysis, preparation, and optimization of statistical sign language machine translation," *Machine Translation*, vol. 26, pp. 325–357, 2012, published online: 18 March 2012.

[6] S. Morrissey, "Data-Driven Machine Translation for Sign Languages," Ph.D. dissertation, Dublin City University, Dublin, Ireland, 2008.

[7] G. Massó and T. Badia, "Dealing with Sign Language Morphemes in Statistical Machine Translation," in *LREC 2010: 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, La Valetta, Malta, 2010, pp. 154–157.

[8] M. Huenerfauth, "Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation," Ph.D. dissertation, University of Pennsylvania, 2006.

[9] M. Filhol, "Combining two synchronisation methods in a linguistic model to describe sign language," in *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, ser. LNCS/LNAI. Springer, 2012, vol. 7206.

[10] G. Morgan and B. Woll, Eds., *Directions in sign language acquisition – Trends in language acquisition research*. Amsterdam: John Benjamins Publishing Company, 2002.

[11] V. Jennings, R. Elliott, R. Kennaway, and J. Glauert, "Requirements For A Signing Avatar," in *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010, pp. 133–136.

[12] C. Baker and C. Padden, "Focusing on the non-manual components of ASL," in *Understanding language through sign language research*, P. Siple, Ed. New York: Academic Press, 1978, pp. 27–57.

[13] G. Coulter, "Raised eyebrows and wrinkled noses: The grammatical function of facial expression in relative clauses and related constructions," in *ASL in a bilingual, bicultural context. Proceedings of the Second National Symposium on Sign Language Research and Teaching*, F. Caccamise and D. Hicks, Eds. Coronado, CA: National Association of the Deaf, 1978, pp. 65–74.

[14] S. Liddell, "Non-manual signals and relative clauses in American Sign Language," in *Understanding language through sign language research*, P. Siple, Ed. New York: Academic Press, 1978, pp. 59–90.

[15] ——, *American Sign Language syntax*. The Hague: Mouton, 1980.

[16] C. Baker-Shenk, "A Micro-Analysis of the Nonmanual Components of Questions in American Sign Language," Ph.D. dissertation, University of California, Berkeley, 1983.

[17] J.S. Reilly and M. McIntire and U. Bellugi, "The acquisition of conditionals in American Sign Language: Grammaticized facial expressions," *Applied Psycholinguistics*, vol. 11, pp. 369–392, 1990.

[18] J. Reilly and U. Bellugi, "Competition on the face: Affect and language in ASL motherese," *Journal of Child Language*, vol. 23, pp. 219–239, 1996.

[19] D. Anderson and J. Reilly, "PAH! the acquisition of adverbials in ASL," *Sign Language & Linguistics*, vol. 1, pp. 3–28, 1998.

[20] R. Wilbur, "Phonological and Prosodic Layering of Nonmanuals in American Sign Language," in *The Signs of Language Revisited*, K. Emmorey and H. Lane, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000, pp. 215–244.

[21] J. Reilly and D. Anderson, "Faces: The acquisition of non-manual morphology in ASL," in *Directions in Sign Language Acquisition*, G. Morgan and B. Woll, Eds. Amsterdam: John Benjamins, 2002, pp. 159–181.

[22] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, pp. 267–373, 2012.

[23] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, July 2010, pp. 504–513.

[5] The same is true for DSGS, but such cases do not appear in the train announcement data used for the experiments reported in this paper.

[24] E. F. T. K. Sang and J. Veenstra, “Representing Text Chunks,” in *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, ser. EACL ’99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 173–179.

[25] L. Roth and S. Clematide, “Tagging Complex Non-Verbal German Chunks with Conditional Random Fields,” in *Proceedings of the 12th Konvens*, Hildesheim, Germany, 2014, pp. 48–57.

[26] D.-K. Kang, A. Silvescu, and V. Honavar, “RNBL-MN: A Recursive Naive Bayes Learner for Sequences Classification,” in *PAKDD’06*, 2006.

[27] A. Savkov, J. Carroll, and J. Cassell, “Chunking Clinical Text Containing Non-Canonical Language,” in *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014)*, Baltimore, Maryland USA, 2014, pp. 77–82.

[28] C. Neidle and C. Vogler, “A new web interface to facilitate access to corpora: Development of the asllrp data access interface,” in *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012*, Istanbul, Turkey, 2012.

[29] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. Lee, *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*, 2nd ed. Cambridge: MIT Press, 2001.