Automatic Generation of German Sign Language Glosses from German Words

Jan Bungeroth and Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University – D-52056 Aachen, Germany
{bungeroth, ney}@informatik.rwth-aachen.de

Abstract. In our paper we present a method for the automatic generation of single German Sign Language glosses from German words. Glosses are often used as a textual description of signs when transcribing Sign Language video data. For a machine translation system from German to German Sign Language we apply glosses as an intermediate notational system. Then the automatic generation from given German words is presented. This novel approach takes the orthographic similarities between glosses and written words into account. The obtained experimental results show the feasibility of our methods for word classes like adverbs, adjectives and verbs with up to 80% correctly generated glosses.

1 Introduction

In the field of automatic translation, significant progress has been made by using statistical methods. This was successfully applied to many language pairs where large amounts of data are available in the form of bilingual corpora. Using statistical machine translation (SMT) for Sign Languages would require such corpora too. Unfortunately only few data is available.

Addressing this data scarceness by using glosses as an intermediate Sign Language notation, we provide a new approach to a Sign Language translation system. The method presented in our paper is the first to examine the automatic generation of glosses for German Sign Language (DGS) from German words. It makes use of German base forms and a small bilingual corpus. We show how the glosses are generated and give results for the different word classes. Our results will show which word classes (e.g. adverbs) perform better than others.

2 Notation

For storing and processing Sign Language, a textual representation of the signs is needed. While there are several notation systems covering different linguistic aspects, we focus on the so called gloss notation. Glosses are widely used for transcribing Sign Language video sequences.

In our work, a gloss is a word describing the content of a sign written with capital letters. Additional markings are used for representing the facial expressions. Unfortunately, no standard convention for glosses has been defined yet. Furthermore, the manual annotation of Sign Language videos is a difficult task, so notation variations within one corpus are often a common problem.

In this work, the gloss notation basically follows the definitions as used in [1]. Additionally, compound nouns are separated with a plus if they are signed separately, and references to locations in signing space, signed with the hands, are given as an X with the location name.

Table 1 shows example glosses representing DGS signs. The glosses, retrieved from DGS video sequences, are given with their English translation.

Table 1. Example glosses for DGS signs and their English translations

$\overset{\text{neg}}{\overline{\text{SITZEN}}}$	GELD+MÜNZEN	X-location
not sitting	money coins	reference to the location X

3 Translation System

A complete Sign Language translation system, capable of generating Sign Language output from spoken input and for generated speech from recognized Sign Language, was proposed in [2].

The system propagates the use of a gloss notation for the corpus-based learning mechanisms. The input sentence (e.g. German) will be translated into glosses which are reordered according to the Sign Language grammar (e.g. DGS grammar). The corresponding animation performed by an avatar, that is a virtual signer, can be looked up in lexicons. Unknown glosses are still useful, as they can be finger-spelled.

4 Corpus

Bilingual Sign Language corpora are still rare, as the consistent annotation of videos is difficult. The available corpora are limited to a few hundred sentences, often taken from different domains. The European Cultural Heritage Online (ECHO) project [3] hosts a number of well annotated, small corpora from various Sign Languages like Swedish Sign Language (SSL), Dutch Sign Language (NGT) and British Sign Language (BSL). Furthermore, ECHO also published guidelines for annotation [4] and suitable software.

For our experiments we rely on a bilingual corpus, from the DESIRE team [5] for DGS and German consisting of 1399 sentences after pre-processing. Table 2 shows the corpus statistics where singletons are words occurring only once. Due to the high number of singletons, this corpus is unsuitable for the immediate training in a SMT system.

Unfortunately, several sentences in the corpus use inconsistent annotation. Also some notations had to be changed. The altered notation will be used for testing the generated glosses later.

5 Gloss Generation

Obtaining a DGS gloss from a given German word is possible because the notation of the DGS sign is described with one or more German words. This similarity

Table 2. DESIRE corpus statistics

	DGS	German
no. of sentence pairs	1399	
no. of running words	5480	8888
no. of distinct words	2531	2081
no. of singleton words	1887	1379

is also dependent on the semantic context of the DGS sentence and it's grammar. We can therefore expect words from some word classes to be generated better than those of others classes. Thus analyzing the word class of the German word is one basic idea of gloss generation.

For this analysis we rely on the commercially available analyzer by Lingsoft[1]. It writes the corresponding morpho-syntactical information of a German sentence to a file.

As an example, we look at the German sentence "Ich mag keine Nudeln." (*I don't like noodles.*). First we extract the base forms and process them for obtaining gloss-like words. That is, special symbols are removed and the obtained words are capitalized. Here, the resulting glosses would be: ICH, MÖGEN, KEIN, NUDEL. We then extract the word classes from this output. In this example that is pronoun (PRON), verb (V), determiner (DET) and noun (S). Note that an ambiguous word can have different interpretations.

Table 3 shows a further example sentence, where the German words are transformed to glosses.

Table 3. Example gloss generation

German	Ich kaufe heute ein neues Auto.
German base forms	ich kaufen heute ein neu Auto
Correct glosses	ICH KAUFEN HEUTE NEU AUTO
Correct DGS	HEUTE NEU AUTO KAUFEN
English	Today I buy a new car.

6 Results

For our experiments we extracted all the base forms of the German sentences in the DESIRE corpus. From these we generated the glosses using the methods described above. The resulting glosses were then compared with the DGS lexicon extracted from the DGS part of the corpus. All generated words were sorted according to their extracted base form. As mentioned in the last section about preprocessing, markings were handled as separate words.

With no further pre-processing we already achieved 55.7% correct matches overall. When looking at the distinct word classes different matching rates were found. Especially adverbs, adjectives and verbs could be generated easily and

[1] http://www.lingsoft.fi

Table 4. Automatic gloss generation for different word categories

	NOUNS	VERBS	ADJ	ADV	PREP	PRON	CONJ	ART
no. of running words	940	924	304	321	248	598	67	275
no. of distinct words	710	210	135	61	35	30	11	10
correct glosses [in %]	52.1	67.1	72.6	81.0	48.6	46.7	0.0	0.0

with a high compliance. Nouns were only generated correctly below average (52.1%). This is explained by a high number of compound nouns that are concatenated differently in DGS than in German. Also synonymous nouns are often used for the DGS transcription, so the generated gloss might be correct but not part of the lexicon. Further investigation on different corpora is necessary for noun generation.

On the other hand, the lack of conjunctions and articles is no surprise as words from these categories are rarely or even never used by signers in DGS. Preposition and pronouns should be handled with care, as those are often used in German, but in DGS they are often substituted by classifier predicates.

7 Summary and Outlook

We described how to generate single Sign Language glosses from given words. This method will be embedded into a complete translation system as described in this paper. The necessary corpus preparation was introduced as well as an overview of the gloss notation.

The generation process itself, where glosses are derived from the base form words, will assist the translation system. From the observed results we conclude to introduce automatic gloss generation for adjectives, adverbs and verbs. It should be possible to alter nouns during preprocessing for obtaining better results on this word class too. This will be addressed on other corpora as our next step towards automatic Sign Language translation.

References

1. C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure.* MIT Press, Cambridge MA, 2000.
2. J. Bungeroth, and H. Ney. Statistical Sign Language Translation. *Workshop on Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 105–108, Lisbon, Portugal, May 2004.
3. O. Crasborn, E. van der Kooij, and J. Mesch. European Cultural Heritage Online (ECHO): Publishing sign language data on the internet. *8th conference on Theoretical Issues in Sign Language Research (TISLR8)*, Barcelona, October 2004.
4. A. Nonhebel, O. Crasborn, and E. van der Kooij. *Sign language transcription conventions for the ECHO project.* ECHO Project, 20 January 2004. Radboud University Nijmegen.
5. DESIRE. *DGS-Phrasensammlung.* Microbooks, Aachen, 1998.