Enriching the Swedish Sign Language Corpus with Part of Speech Tags Using Joint Bayesian Word Alignment and Annotation Transfer

Robert Östling, Carl Börstell, Lars Wallin
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
{robert,calle,wallin}@ling.su.se

Abstract

We have used a novel Bayesian model of joint word alignment and part of speech (PoS) annotation transfer to enrich the Swedish Sign Language Corpus with PoS tags. The annotations were then hand-corrected in order to both improve annotation quality for the corpus, and allow the empirical evaluation presented herein.

1 Introduction

While Swedish Sign Language (SSL) is a recognized language of Sweden, there are no NLP tools available for it. We used a novel method for part of speech (PoS) annotation transfer to create the first automatically PoS-tagged corpus of any sign language, by exploiting the existing translation of the corpus into written Swedish. We then manually corrected the resulting annotations, in order to provide high-quality data that could be used for e.g. future supervised PoS taggers.

2 Parts of Speech in Sign Languages

In early sign language (SL) linguistics, Supalla and Newport (1978) observed that consistent phonological patterns can distinguish PoS in phonologically and semantically related noun–verb pairs in American SL: nouns more often being restrained and repeated; verbs having a continuous articulation. A number of studies have similarly investigated nouns and verbs in a variety of SLs—e.g. Johnston (2001) for Australian SL; Hunger (2006) for Austrian SL; Kimmelman (2009) for Russian SL; and Tkachman and Sandler (2013) for Al-Sayyid Bedouin SL and Israeli SL—finding that e.g. manner, repetition, duration, size and mouthing[1] can differentiate nouns from verbs. However, extensive research of PoS in SLs is generally lacking (Schwager and Zeshan, 2008).

Turning to SSL, one study looked at morphosyntactic constraints differentiating e.g. verbs and adjectives (Bergman, 1983), and a recent overview of the linguistic structure of SSL lists 8 PoS (Ahlgren and Bergman, 2006)—based mainly on semantic and syntactic criteria identified in previous research—which is the set used in this work.

There are currently a number of ongoing SL corpus projects around the world, but few of them have extensive annotations for e.g. grammatical categories. One exception is the Auslan Corpus (Johnston, 2010), which features annotations of "grammatical classes" that to a large extent correspond to functional PoS (Johnston, 2014). However, the annotations in all of these SL corpora are done manually, which is rather time-consuming.

3 The Swedish Sign Language Corpus

The Swedish Sign Language Corpus (SSLC) (Mesch et al., 2012; Mesch et al., 2014) contains 25 hrs of partially transcribed, conversational data from 42 different signers of SSL. Its annotations mainly consist of a gloss for each sign (see 4.3), and a translation into written Swedish. The version used in our evaluations contains 24 976 SSL tokens, not sentence-segmented, and 41 910 Swedish tokens divided into 3 522 sentences.

Segmenting SSL data into sentences or utterances is no trivial task (Börstell et al., 2014), and there is currently no such segmentation in the SSLC. In order to use sentence-based word alignment models, we follow Sjons (2013) in using the Swedish sentences as a basis for segmentation.

[1] *Mouthing* refers to a mouth movement imitating that of a spoken word, i.e. as if silently articulating a word.

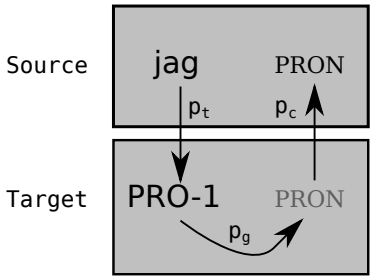

Figure 1: Circular generation model for joint word alignment and part of speech transfer, where Swedish *jag* 'I' is aligned to the SSLC gloss PRO-1. All variables are observed except the alignment and the target-side PoS tag (in grey).

4 Method

Since SSLC contains translations of each utterance into written Swedish, it is in effect a parallel corpus of Swedish and SSL. It has long been recognized that parallel corpora are useful for transfering annotation between languages (Yarowsky et al., 2001), and so our goal is to automatically transfer part of speech (PoS) tags from the Swedish translations to the SSL glosses. Given the relatively small size of the corpus, and the significant differences between the two languages, the error rate is expected to be high enough to warrant a final stage of manual correction.

4.1 Bayesian word alignment

Most previous research on word alignment has been building on the IBM models Brown et al. (1993) using Expectation-Maximization (EM) for inference. Recently, Bayesian models have been proposed as an alternative, offering a theoretically well-founded way of introducing soft constraints that encourage linguistically plausible solutions (DeNero et al., 2008; Mermer and Saraçlar, 2011; Riley and Gildea, 2012; Gal and Blunsom, 2013).

4.2 Alignment and part of speech transfer

Several authors, starting from Brown et al. (1993), have used word classes to aid word alignment in various manners. Toutanova et al. (2002) show that if both parts of a bitext are annotated with PoS tags, alignment accuracy can be improved simply by using a tag translation model $p(t_f|t_e)$ in addition to the word translation model $p(f|e)$.

Given that PoS tags improves the quality of word alignments, and that word alignments can be used to transfer PoS tags from one language to another, we use a model that jointly learns PoS tags and word alignment.

Figure 1 illustrates our *circular generation model*, so termed because a source language token is assumed to generate a target language token (through $p_t(f|e) = \theta_{te}$), which then generates its corresponding PoS tag (through $p_g(t_f|f) = \theta_{gf}$), which finally generates a PoS tag for the source language token (through $p_c(t_e|t_f) = \theta ct_f$).

We assume categorical distributions with Dirichlet priors:

$$\theta_t \sim \mathrm{Dir}(\alpha_t); \theta_g \sim \mathrm{Dir}(\alpha_g); \theta_c \sim \mathrm{Dir}(\alpha_c)$$

These priors are symmetrical, except for α_c which is biased towards consistency between the tagsets. In our evaluations, we use $\alpha_c = 1000$ for consistent tag pairs, $\alpha_c = 1$ for others.

Inference in this model is performed using collapsed Gibbs sampling, where the alignment variable a_j (which links target token f_j to source token e_i) is sampled jointly with the target PoS tag t_{f_j}. The sampling equation is similar to those used by Mermer and Saraçlar (2011) and Gal and Blunsom (2013), with extra factors for the PoS tag dependencies. We also use the HMM-based word order model of Vogel et al. (1996), but as this does not directly interact with the PoS tags, we exclude it from Figure 1 for clarity.

From this model we obtain SSL part of speech tags in one of two different ways: directly from the model (the t_{f_j} variables), or by direct projection through the final alignment variables a_j. In both cases the sampling procedure is identical, the difference lies only in how the final PoS tags are read out.

4.3 Data processing

SSLC is annotated using the ELAN software (Wittenburg et al., 2006). Annotations are arranged into *tiers*, each containing a sequence of annotations with time slots. For the present study, two types of tiers are of interest: the signs of the dominant hand and the Swedish sentences. Signs are transcribed using glosses, whose names are most often derived from a corresponding word or expression in Swedish. Each gloss may also have a number of properties marked, such as which hand it was performed with, whether it was reduplicated, interrupted, and so on. The annotation conventions are described in further detail by Wallin et al. (2014).

The first step of processing is to group SSL glosses according to which Swedish sentence they overlap most with. Second, glosses with certain marks are removed:

- Interrupted signs (marked `@&`).
- Gestures (marked `@g`).
- Incomprehensible signs (transcribed as XXX).

Finally, some marks are simply stripped from glosses, since they are not considered important to the current task.

- Signs performed with the non-dominant hand (marked `@nh`).
- Signs held with the non-dominant hand during production of signs with the dominant hand. The gloss of the held sign (following a <> symbol) is removed.
- Signs where the annotator is uncertain about which sign was used (marked `@xxx`) or whether the correct gloss was chosen (marked `@zzz`).

In all, this is nearly identical to the procedure used by Sjons (2013, p. 14). Example 1 illustrates the output of the processing, with English glosses and translation added.

(1) STÄMMA OCKSÅ PRO-1 PERF BARN
be.correct also 1 PRF children
BRUKA SE SAGAˆTRÄD PRO>*närv*
usually watch Sagoträdet 2
'I also have children—do you watch Sagoträdet?'

The Swedish translations were tokenized and PoS-tagged with Stagger (Östling, 2013), trained on the Stockholm-Umeå Corpus (SUC) (Ejerhed et al., 1992; Källgren, 2006) and the Stockholm Internet Corpus (SIC).[2]

4.4 Evaluation data

At the outset of the project, two annotators manually assigned PoS tags to the 371 most frequent sign glosses in the corpus. This was used for initial annotation transfer evaluations, and when the methods reached a certain level of maturity the remaining gloss types were automatically annotated, and the resulting list of 3 466 glosses manually corrected. Thus the initial goal of using annotation transfer to facilitate the PoS annotation was achieved, since all of the currently transcribed SSLC data now has manually verified annotations.

[2] http://www.ling.su.se/sic

PoS	SSLC	SUC
Pronoun	PN	DT, HD, HP, HS, PS, PN
Noun	NN	NN, PM, UO
Verb	VB	PC, VB
Adverb	AB	AB, HA, IE, IN, PL
Numeral	RG	RG, RO
Adjective	JJ	JJ
Preposition	PP	PP
Conjunction	KN	KN, SN

Table 1: PoS tags in the SSLC, and their counterparts in SUC.

In order to evaluate the performance of the annotation transfer algorithms, we use this final set of 3 466 annotated types as a gold standard.

4.5 Tag set conversion

As previously mentioned, we use the eight PoS categories suggested by Ahlgren and Bergman (2006) for SSL. The Swedish side is tagged using the SUC tagset, whose core consists of 22 tags (Källgren, 2006, p. 20). For direct tag projection and the tag translation priors in the circular generation model, the SUC tags are translated as in Table 1.

4.6 Task-specific constraints

SSLC contains various annotations that are useful for part of speech tagging. First of all, a few parts of speech are already apparent from the annotation: proper nouns (marked `@en`), pronouns (begin with PRO- or POSS-), and classifier constructions (marked `@p`, considered to be verbs). Second, the choice of gloss (in Swedish) correlates strongly with the SSL part of speech. In Example 1, for instance, the part of speech of most signs can be correctly guessed from the gloss alone, although sometimes this is not possible due ambiguity (e.g. *stämma* is ambiguous between noun and verb) or because the gloss name is not a Swedish word (e.g. PERF). We exploit this correspondence by requiring glosses to be tagged consistently with the SALDO morphological lexicon (Borin and Forsberg, 2009). That is, if the SALDO lexicon says that the name of a gloss is a Swedish word form with an unambiguous word class, this word class will always be assumed for the gloss. As can be seen from the baseline row in Table 2,

	Types		Tokens	
	Project	**Model**	**Project**	**Model**
baseline	$58.4 \pm 0.5\%$	$12.2 \pm 0.6\%$	$75.3 \pm 0.7\%$	$10.8 \pm 3.6\%$
constraints	$58.4 \pm 0.4\%$	$60.7 \pm 0.4\%$	$75.1 \pm 0.8\%$	$58.1 \pm 1.0\%$
circular	$64.7 \pm 0.5\%$	$68.3 \pm 0.4\%$	$77.4 \pm 0.8\%$	$77.6 \pm 0.7\%$
circular + bigrams	$64.8 \pm 0.3\%$	$68.4 \pm 0.3\%$	$77.3 \pm 0.7\%$	$77.6 \pm 0.7\%$
circular + constraints	$69.1 \pm 0.4\%$	$77.1 \pm 0.3\%$	$79.7 \pm 0.6\%$	$78.7 \pm 0.6\%$

Table 2: Token-level PoS tagging accuracy, using direct projection from the final alignment (projection) or for the joint models, the sampled PoS tag variables t_{f_j} (model). Note that in the former case, the PoS tag variables are ignored except during the alignment process. Figures given are averages $\pm$ standard deviation estimated over 64 randomly initialized evaluations for each configuration.

these constraints alone reach a fairly high level of accuracy.

5 Results

Table 2 shows the per-type and per-token accuracy for a number of different configurations, using both direct projection (*project*) and the sampled t_{f_j} PoS tag variables (*model*). While the model itself assigns tags on the token level, when obtaining the figures in Table 2 we ensure type-level tagging by assigning each token the majority tag of its type. This is also done for the *tokens* columns. Since the SSLC annotation does not contain glosses with ambiguous part of speech, using only token-level would introduce unnecessary noise.

The *baseline* model performs word alignment only, assigning random values to the PoS tag variables. Projection accuracy is fair (indicating that the word alignments are acceptable), but the much lower scores in the *model* columns represent an entirely random baseline. When the constraints described in Section 4.6 are enforced during sampling, the *model* scores increase as expected. Without coupling between the word alignment and PoS tags, the projected tags are not changed.

Using the circular generation model, the coupling between PoS tags and word alignment leads to better accuracy, as expected. This is also the case when using direct projection, indicating that the joint model increases word alignment quality (which we are unable to evaluate directly, lacking a gold standard word alignment).

Adding bigram dependencies between SSLC PoS tags allows the model to use (monolingual) contextual information, but this does not seem to affect the accuracy. The probable reason for this is that the coupling between Swedish and SSLC PoS tags is quite strong, and the contextual information can not significantly affect the final tagging.

The best results were obtained by using the circular model in conjunction with enforcing the task-specific constraints. The improvement when adding the constraints is particularly large on the type level, because this allows reasonable guesses for many of the rare types where there is not enough data for reliable word alignments. This is important for our goal, since we actually want type-based part of speech assignments.

The 77% accurate type-level tag list was manually corrected by two annotators (reaching consensus) in order to obtain the final tagging of the corpus, which is also used as the gold standard for the evaluation described in this section.

6 Conclusions

We have shown that annotation transfer, performed in a Bayesian model of joint word alignment and part of speech transfer, can be a useful tool when annotating a sign language corpus with parts of speech.

It should be stressed that our conclusions apply to this particular data set, which is rather untypical for annotation projection tasks, especially due to its limited size. Work carried out in parallel with the present study indicates that for longer parallel texts with better translations, other models may be more accurate. Since the primary aim of this study was to investigate methods for annotating the SSLC, rather than exploring annotation transfer in general, we are not currently concerned with other data.

References

Inger Ahlgren and Brita Bergman. 2006. Det svenska teckenspråket. In *Teckenspråk och teckenspråkiga: kunskaps- och forskningsöversikt*, volume 2006:29 of *Statens offentliga utredningar (SoU)*, pages 11–70. Ministry of Health and Social Affairs, March.

Brita Bergman. 1983. Verbs and adjectives: morphological processes in Swedish Sign Language. In Jim Kyle and Bencie Woll, editors, *Language in sign: An international perspective on sign language*, pages 3–9, London. Croom Helm.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *NODALIDA 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, pages 7–12, Odense, Denmark.

Carl Börstell, Johanna Mesch, and Lars Wallin. 2014. Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages*, pages 7–10, Reykjavík, Iceland. ELRA.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October. Association for Computational Linguistics.

E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project. Technical report, Department of Linguistics, University of Umeå.

Yarin Gal and Phil Blunsom. 2013. A systematic Bayesian treatment of the IBM alignment models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barbara Hunger. 2006. Noun/verb pairs in Austrian Sign Language (ÖGS). *Sign Language & Linguistics*, 9(1/2):71–94, January.

Trevor Johnston. 2001. Nouns and verbs in Australian sign language: An open and shut case? *Journal of Deaf Studies and Deaf Education*, 6(4):235–57, January.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2014. Auslan corpus annotation guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University.

Vadim Kimmelman. 2009. Parts of speech in Russian Sign Language: The role of iconicity and economy. *Sign Language & Linguistics*, 12(2):161–186.

Gunnel Källgren, 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University, December. Sofia Gustafson-Capková and Britt Hartmann (eds.).

Coşkun Mermer and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 182–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. Sign language resources in Sweden: Dictionary and corpus. In *Proceedings 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Language Resources and Evaluation Conference (LREC), pages 127–130, Istanbul, Turkey.

Johanna Mesch, Maya Rohdell, and Lars Wallin. 2014. Annoterade filer för svensk teckenspråkskorpus. Version 2. `http://www.ling.su.se`.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for Swedish. *North European Journal of Language Technology*, 3:1–18.

Darcey Riley and Daniel Gildea. 2012. Improving the IBM alignment models using variational Bayes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 306–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

Waldemar Schwager and Ulrike Zeshan. 2008. Word classes in sign languages: Criteria and classifications. *Studies in Language*, 32(3):509–545, September.

Johan Sjons. 2013. Automatic induction of word classes in Swedish Sign Language. Master's thesis, Stockholm University.

Ted Supalla and Elissa L. Newport. 1978. How many seats in a chair?: The derivation of nouns and verbs in American Sign Language. In Patricia Siple, editor, *Understanding language through sign language research*, chapter 4, pages 91–132. Academic Press, New York, NY.

Oksana Tkachman and Wendy Sandler. 2013. The noun-verb distinction in two young sign languages. *Gesture*, 13(3):253–286.

Kristina Toutanova, H. Tolga Ilhan, and Christopher Manning. 2002. Extensions to HMM-based statistical word alignment models. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 87–94.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lars Wallin, Johanna Mesch, and Anna-Lena Nilsson. 2014. Transkriptionskonventioner för teckenspråkstexter (version 5). Technical report, Sign Language, Department of Linguistics, Stockholm University.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. `http://tla.mpi.nl/tools/tla-tools/elan/`.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.