Automatic interlinear glossing as two-level sequence classification

Tanja Samardžić
Corpus Lab
URPP Language and Space
University of Zurich

Robert Schikowski
Department of
Comparative Linguistics
University of Zurich

Sabine Stoll
Department of
Comparative Lingistics
University of Zurich

tanja.samardzic | robert.schikowski | sabine.stoll@uzh.ch

Abstract

Interlinear glossing is a type of annotation of morphosyntactic categories and cross-linguistic lexical correspondences that allows linguists to analyse sentences in languages that they do not necessarily speak. Automatising this annotation is necessary in order to provide glossed corpora big enough to be used for quantitative studies. In this paper, we present experiments on the automatic glossing of Chintang. We decompose the task of glossing into steps suitable for statistical processing. We first perform grammatical glossing as standard supervised part-of-speech tagging. We then add lexical glosses from a stand-off dictionary applying context disambiguation in a similar way to word lemmatisation. We obtain the highest accuracy score of 96% for grammatical and 94% for lexical glossing.

1 Introduction

The annotation type known as interlinear glossing allows linguists to describe the morphosyntactic makeup of words concisely and language-independently. While glosses as a linguistic metalanguage have a long tradition, systematic standards for interlinear glossing have only developed relatively recently – cf. e.g. the Leipzig glossing rules.[1]

An example for an interlinear gloss is shown in (1), which is an Ewe serial verb construction taken from (Collins, 1997) with glosses in boldface. The combination of segmentation with English metalanguage labels for both lexical and grammatical segments allows linguists to observe how exactly the Ewe serial verb construction differs from the corresponding English construction.

(1) *Kofi tso ati-ɛ fo Yao (yi).*
Kofi take stick-DEF hit Yao P
Kofi took the stick and hit Yao with it.

With the development of annotated linguistic corpora of various languages, glosses are starting to be used in a new way. Traditionally, only individual sentences or small text collections were glossed to illustrate examples. Nowadays glosses are systematically added to large corpora in order to provide structural information necessary for quantitative cross-linguistic research.

Despite their great value for linguistic research, glossed corpora often remain rather small. The main reason for this is the fact that glossing requires a high level of linguistic expertise and is currently performed manually by trained experts. This practice makes the creation of glossed corpora extremely time-consuming and expensive. In order to obtain glossed corpora large enough for reliable quantitative analysis, the process of glossing needs to be automatised.

In this paper, we present a series of experiments performed with this precise aim.[2] We divide the traditional glossing procedure into several steps and define an automatic processing pipeline, which consists of some standard and some custom natural language processing tasks. The data we use for our experiments come from the Chintang Language Corpus (Bickel et al., 2004 2015), an exceptionally large glossed corpus, which has been developed since 2004 and is presently hosted at the Department of Comparative Linguistics at the University of Zurich.[3]

2 Related work

Data for comprehensive linguistic research need to be collected in a wide range of languages. Glosses

[1] Available at https://www.eva.mpg.de/lingua/resources/glossing-rules.php

[2] This work is partially supported by the S3IT computing facilities.

[3] http://www.clrp.uzh.ch

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 68–72,
Beijing, China, July 30, 2015.

are especially important for research in under-resourced languages, the analysis of which requires more detailed information than it is the case with well documented and processed languages.

Approaches to under-resourced languages include developing tools to support manual rule crafting and deep rule-based analysis (Bender et al., 2014; Snoek et al., 2014), data collection by experts (Ulinski et al., 2014; Hanke and Bird, 2013) and through crowd-sourcing (Bird et al., 2014; Dunham et al., 2014), automatic translation and cross-linguistic projection using parallel corpora (Yarowsky et al., 2001; Scherrer, 2014; Scherrer and Sagot, 2014; Aepli et al., 2014), and part-of-speech tagging (Garrette and Baldridge, 2013).

These tasks target different representations, but the resources that they produce are not suitable for corpus-based quantitative linguistic research. Our approach to automatic linguistic glossing is intended to fill this gap. Like Garrette and Baldridge (2013), we learn our target representation from a relatively small sample of manually developed resources in the target language. However, we tackle a harder task and rely on more resources, which are becoming increasingly available through work on language documentation.

3 The structure of the Chintang corpus and glossing strategy

The corpus consists of about 290 hours (1,232,161 words) of video materials transcribed in broad IPA. 214 hours (955,025 words) have been translated to English and Nepali by native research assistants and glossed by trained non-native student assistants. The primary data are MPG-1 videos and WAV audio files. Morphological Transcriptions and translations were done in Transcriber and ELAN, glossing in Toolbox.

The basic unit of a Toolbox text is the record, which in oral corpora usually corresponds to one utterance. Records are separated from each other by double newlines. Each record may contain several tiers, each of which is coded as a line ended by a single newline. Within each tier, both words and morphemes are separated by whitespace. Free and bound morphemes are distinguished by adding a corresponding separator (prefixes end and suffixes begin with a "-"). Elements are implicitly aligned across tiers based on their position on a tier (e.g. the fifth element on the segment tier corresponds to the fifth element on the gloss tier).

record ID	rabbit.047			
transcription	mande		aba	katha
segmentation	mand	-e	abo	katha
glosses	be.over	-ind.pst	now	story
language	C	-C	C/N	N
lexicon ID	281	-1234	596	4505
PoS	vi	-gm	adv	n
English	Now the story is over.			

Table 1: Example for record structure in the Chintang Language Corpus

lexeme	mand
ID	281
variant	mai
PoS	vi
valency	S-NOM(1) V-s(S)
English gloss	be.finished; be.over; be.used.up
language	C

Table 2: Example for entry structure in the Chintang dictionary

Table 1 shows a simplified example of an analysed record. Beside segmentation and interlinear glosses, the analysis also includes POS tags, language labels, and lexical IDs for every morpheme.

The language labels are needed because mixing with other languages (Nepali and Bantawa) is frequent in Chintang (Stoll et al., 2015).

The lexical IDs provide a unique link to the entries of an electronic lexicon of Chintang, which contains rich information both on free and on dependent morphemes. A simplified example of a lexicon entry is given in Table 2.

Lexical IDs ensure good communication between the corpus and the lexicon, allowing for queries involving both resources at the same time (e.g. combining valency or etymology information from the lexicon with corpus counts) as well as systematic updates and synchronisation of both resources.

4 Automatic glossing pipeline

As shown in the previous section, linguistic annotation of the Chintang corpus consists of word segmentation and proper glossing.

Word segmentation can be seen as a preprocessing step that creates basic units of analysis

to which glosses are assigned. Once the words are segmented into morphemes that encode either lexical content or grammatical categories, assigning glosses to the word segments reduces to a one-to-one mapping: each segment in a sentence is assigned exactly one gloss and vice versa.

We take manual segmentation as input and learn automatically the mappings between segments and glosses. This mapping includes referring to the corresponding lexicon and importing the information from lexical entries.

A simple way to learn this mapping would be to treat glossing as word-by-word translation from original text to an artificial language that consists of words and grammatical tags. Word order in the artificial language would be exactly the same as in the original text, so that the task would reduce to learning segment translation probabilities.

The main disadvantage of the translation-based approach is that it requires large corpora for training. This approach does not generalise beyond examples seen in the training set, which is why good coverage of a new text can be obtained only if the model is trained on a large corpus. In addition to this, the translation model would need to be complemented by a language model to account for context dependencies. However, large glossed corpora are almost never available for under-resourced languages, as discussed above.

Another possible approach is to treat glosses as a special kind of part-of-speech tags. The main obstacle for this approach is the fact that glosses contain lexical items (lexical glosses). Including lexical tags would result in a tag set too big to be learnt by standard part-of-speech tagging models.

We thus apply a two-level tagging approach where we first learn grammatical tags without lexical items in a standard supervised part-of-speech tagging setting. We then add lexical items from the lexicon using the sequences of grammatical tags for disambiguation. In the remainder of this section, we describe the two procedures in more detail and experiments performed to evaluate them.

4.1 Grammatical annotation as PoS tagging

To separate grammatical from lexical glossing, we merge two tiers of the original annotation. This is done by replacing lexical items in the gloss tier with their corresponding part-of-speech tags. In the case of grammatical items, we keep the original gloss. This results in a representation illustrated in (2), where the lexical glosses *be.over*, *now*, and *story* are replaced by their corresponding part-of-speech tags *vi*, *adv*, and *n*.

(2)

mand	*-e*	*abo*	*katha*
vi	-ind.pst	adv	n

In this way, we obtain a corpus annotated with 233 distinct labels that describe relevant morphosyntactic categories in Chintang.

We then split the corpus into a train and a test set and apply a standard supervised part-of-speech tagging. We train and test a general-purpose state-of-the-art statistical tagger (Gesmundo, 2011; Gesmundo and Samardžić, 2012).

To assess how the quantity of training data influences the performance of the tagger, we run the experiment several times using increasing amounts of data for training. The results of the tagging experiments are presented in Figure 1.

4.2 Lexical annotation as lemmatisation

To recover the original glosses, we replace part-of-speech tags of words with lexical content by their corresponding English lemmas. English lemmas are associated to their corresponding Chintang segments in the lexicon, where each entry is identified with a unique numerical code (lexion ID). The task of inserting lexical glosses back is thus reduced to the task of finding the correct lexicon ID for each word segment in a sentence. We perform this in two steps.

In the first step, we search the lexicon to find all possible IDs for a given pair consisting of a segment and its grammatical tag assigned by the tagger. We select all entries where the given word appears as the main entry or as a variant, and the given grammatical tag appears either as the gloss or as the part-of-speech tag (see Table 2). Even though we look up word segments disambiguated for their grammatical category, approximately 15% of the pairs remain ambiguous in the sense that multiple possible lexicon IDs are assigned.

In the second step, we select a single ID through a disambiguation procedure that takes into account the previous context of the segment. This step is similar to the procedures used in the task of lemmatisation. We represent the previous context with a sequence of two grammatical tags assigned by the tagger to the preceding segments, t_{-2}, and t_{-1}. We estimate the probability of each of the possible

lexicon IDs (id_0) given the previous two tags. We then select the most probable ID, as shown in (3).

$$\begin{aligned} id_0* &= arg\,max_{id_0} p(id_0|t_{-2}, t_{-1}) \\ &= arg\,max_{id_0} \frac{p(t_{-2}, t_{-1}, id_0)}{p(t_{-2}, t_{-1})} \end{aligned} \quad (3)$$

In the cases where the trigram-based estimation is not possible due to zero counts, we apply a three-step back-off strategy. If counts can be collected, we find the most probable ID given only one previous tag, as shown is (4). Otherwise, we select the most likely ID without the context information. Finally, if there are no corpus counts, we select one of the possible IDs randomly.

$$\begin{aligned} id_0* &= arg\,max_{id_0} p(id_0|t_{-1}) \\ &= arg\,max_{id_0} \frac{p(t_{-1}, id_0)}{p(t_{-1})} \end{aligned} \quad (4)$$

We evaluate lexical annotation in the same settings as in the case of part-of-speech tagging. The results are shown in Figure 1.

4.3 Results and discussion

Figure 1 shows the performance on the two tasks using different corpus sizes for training. In each run, we increase the length of the training set by approximately 50,000 tokens, keeping the test set constant (around 200,000 tokens).

The accuracy of part-of speech tagging obtained with the first set (50,000 tokens) is 90%. It increases by 1% with every increase till the size of 200,000. The increase after this point is much slower, reaching the best result of 96% accuracy using the full training set of around 800,000 tokens.

The performance curve is a little different for the task of lexical annotation. The accuracy is already 92% when the disambiguation model is trained on the smallest set. It reaches the maximal score of 94% with the training set of 200,000 tokens.

These results show that dividing glossing into two sequence classification tasks allows us to optimise manual work in developing new resources. A relatively small annotated corpus is used to model sequences of highly frequent items (grammatical words and their tags). Sparse but less ambiguous lexical items are glossed using a lexicon, ensuring good coverage. In this framework, new segments are addressed in two ways. Grammatical tags are assigned to new words based on the generalisations learnt by the tagger. Lexical tags are

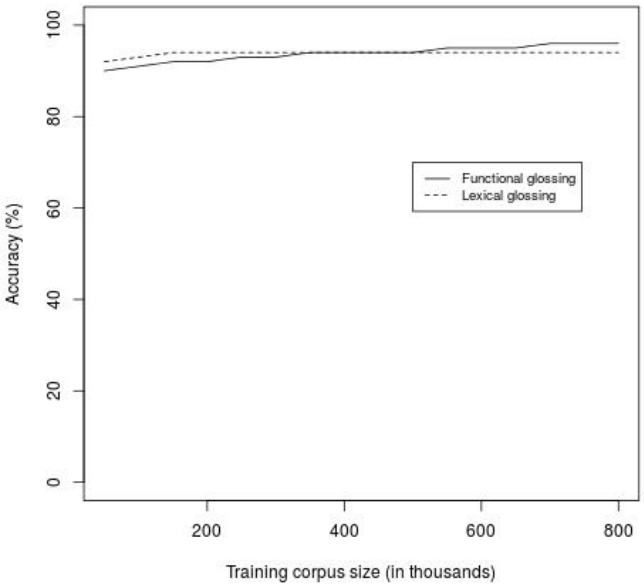

Figure 1: Performance on two glossing subtasks using increasing sizes of the train set.

expected to be covered by the lexicon. Items that are not covered need to be manually added to the lexicon, but they are then automatically applied in glossing.

A number of mismatches between predicted labels and the gold standard are caused by inconsistencies in the gold standard due to the changes in the label set over time. While we count all the mismatches as errors, an inspection of the output of automatic processing can be used to improve annotation consistency.

5 Conclusion and future work

We have shown in this paper how statistical natural language processing techniques can be adapted to the task of interlinear glossing, with the quality of the processing high enough to replace manual annotation. While an annotated sample corpus in the target language is needed to train the statistical models, we show that the initial training set can be relatively small, of the size of some existing glossed corpora.

A fully automatised glossing procedure would have to include an approach to word segmentation, which is not addressed here. We identify this task as the first step in our future work.

References

Noemi Aepli, Ruprecht von Waldenfels, and Tanja Samardžić. 2014. Part-of-speech tag disambiguation by cross-linguistic majority vote. In *First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from igt: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Balthasar Bickel, Sabine Stoll, Martin Gaenszle, Novel Kishor Rai, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Netra Paudyal, Judith Pettigrew, Ichchha P. Rai, and Manoj Rai. 2004-2015. Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children. DOBES Archive, `http://www.mpi.nl/DOBES`.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Chris Collins. 1997. Argument sharing in serial verb constructions. *Linguistic Inquiry*, 28:461–497.

Joel Dunham, Gina Cook, and Joshua Horner. 2014. Lingsync & the online linguistic database: New models for the collection and management of data for language communities, linguists and language learners. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 24–33, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147. Association for Computational Linguistics.

Andrea Gesmundo and Tanja Samardžić. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea, July. Association for Computational Linguistics.

Andrea Gesmundo. 2011. Bidirectional Sequence Classification for Tagging Tasks with Guided Learning. In *Proceedings of TALN 2011*.

R. Florian Hanke and Steven Bird. 2013. Large-scale text collection for unwritten languages. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1134–1138. Asian Federation of Natural Language Processing.

Yves Scherrer and Benoît Sagot. 2014. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Yves Scherrer. 2014. Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland. Association for Computational Linguistics.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of plains cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Sabine Stoll, Taras Zakharko, Steven Moran, Robert Schikowski, and Balthasar Bickel. 2015. Syntactic mixing across generations in an environment of community-wide bilingualism. *Frontiers in Psychology*, 6(82).

Morgan Ulinski, Anusha Balakrishnan, Daniel Bauer, Bob Coyne, Julia Hirschberg, and Owen Rambow. 2014. Documenting endangered languages with the wordseye linguistics tool. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 6–14, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st international conference Human Language Technology*, pages 161–168, San Diego, CA. Association for Computational Linguistics.