Evaluating Automation Strategies in Language Documentation

Alexis Palmer, Taesun Moon, and Jason Baldridge
Department of Linguistics
The University of Texas at Austin
Austin, TX 78712
{alexispalmer,tsmoon,jbaldrid}@mail.utexas.edu

Abstract

This paper presents pilot work integrating machine labeling and active learning with human annotation of data for the language documentation task of creating interlinearized gloss text (IGT) for the Mayan language Uspanteko. The practical goal is to produce a totally annotated corpus that is as accurate as possible given limited time for manual annotation. We describe ongoing pilot studies which examine the influence of three main factors on reducing the time spent to annotate IGT: suggestions from a machine labeler, sample selection methods, and annotator expertise.

1 Introduction

Languages are dying at the rate of two each month. By the end of this century, half of the approximately 6000 extant spoken languages will cease to be transmitted effectively from one generation of speakers to the next (Crystal, 2000). Under this immense time pressure, documentary linguists seek to preserve a record of endangered languages while there are still communities of speakers to work with. Many language documentation projects target languages about which our general linguistic knowledge is nonexistent or much less than for more widely-spoken languages. The vast majority of these are individual or small-group endeavors on small budgets with little or no institutional guidance by the greater documentary linguistic community. The focus in such projects is often first on collection of data (documentation), with a following stage of linguistic analysis and description. A key part of the analysis process, detailed linguistic annotation of the recorded texts, is a time-consuming and tedious task usually occurring late in the project, if it occurs at all.

Text annotation typically involves producing interlinearized glossed text (IGT), labeling for morphology, parts-of-speech, etc., which greatly facilitates further exploration and analysis of the language. The following is IGT for the phrase *xelch li* from the Mayan language Uspanteko:[1]

(1) x- el -ch li
COM- salir -DIR DEM

Spanish: 'Salio entonces.' **English**:'Then he left.'

The levels of analysis include morpheme segmentation, transliteration of stems, and labeling of stems and morphemes with tags, some corresponding to parts-of-speech and others to semantic distinctions.

There is no single standard format for IGT. The IGT systems developed by documentation projects tend to be idiosyncratic: they may be linguistically well-motivated and intuitive, but they are unlikely to be compatible or interchangeable with systems developed by other projects. They may lack internal consistency as well. Nonetheless, IGT in a readily accessible format is an important resource that can be used fruitfully by linguists to examine hypotheses on novel data (e.g. Xia and Lewis (2007; 2008), Lewis and Xia (2008)). Furthermore, it can be used by educators and language activists to create curriculum material for mother language education and promote the survival of the language.

Despite the urgent need for such resources, IGT annotations are time consuming to create entirely by hand, and both human and financial resources are extremely limited in this domain. Thus, language

[1] KEY: COM=completive aspect, DEM=demonstrative, DIR=directional

Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, pages 36–44,
Boulder, Colorado, June 2009.

documentation presents an interesting test case and an ideal context for use of machine labeling and active learning. This paper describes a series of experiments designed to assess this promise in a realistic documentation context: creation of IGT for the Mayan language Uspanteko. We systematically compare varying degrees of machine involvement in the development of IGT, from minimally involved situations where examples for tagging are selected sequentially to active learning situations where the machine learner selects samples for human tagging and suggests labels. We also discuss the challenges faced by linguists in having to learn, transcribe, analyze, and annotate a language almost simultaneously and discuss whether machine involvement reduces or compounds those challenges.

In the experiments, two documentary linguists annotate IGT for Uspanteko texts using different levels of support from a machine learned classifier. We consider the interaction of three main conditions: (1) sequential, random, or uncertainty sampling for requesting labels from an annotator, (2) suggestions or no suggestions from a machine labeler, and (3) expert versus non-expert annotator. All annotator decisions are timed, enabling the actual time cost of annotation to be measured within the context of each condition. This paper describes the Uspanteko data set we adapted for the experiments, expands on the choices described above, and reports on preliminary results from our ongoing annotation experiments.

2 Data: Uspanteko IGT

This section describes the Uspanteko corpus used for the experiments, our clean-up of the corpus, and the specific task—labeling part-of-speech and gloss tags—addressed by the experiments.

2.1 OKMA Uspanteko corpus

Our primary dataset is a corpus of texts (Pixabaj et al., 2007) in the Mayan language Uspanteko that were collected, transcribed, translated (into Spanish) and annotated as part of the OKMA language documentation project.[2] Uspanteko, a member of the K'ichee' branch of the Mayan language family, is spoken by approximately 1320 people in central Guatemala (Richards, 2003).

[2] http://www.okma.org

The corpus contains 67 texts, 32 of them glossed. Four textual genres are represented in the glossed portion of the corpus: oral histories (five texts) usually have to do with the history of the village and the community, personal experience texts (five texts) recount events from the lives of individual people in the community, and stories (twenty texts) are primarily folk stories and children's stories. The corpus also contains one recipe and one advice text in which a speaker discusses what the community should be doing to better preserve and protect the environment.

The transcriptions are based on spoken data, with attendant dysfluencies, repetitions, false starts, and incomplete sentences. Of the 284,455 words, 74,298 are segmented and glossed. This is a small dataset by computational linguistics standards but rather large for a documentation project.

2.2 Interlinearized Glossed Text

Once recordings have been made, the next tasks are typically to produce translations and transcription of the audio. Transcription is a complex and difficult process, often involving the development of an orthography for the language in parallel. The product of the transcription is raw text like the Uspanteko sample shown below (text 068, clauses 283-287):

```
  Non li in yolow rk'il kita'
tinch'ab'ex laj inyolj iin, si no ke
laj yolj jqaaj tinch'ab'ej i non qe li
xk'am rib' chuwe, non qe li lajori
non li iin yolow rk'ilaq.[3]
```

Working with the transcription, the translation, and any previously-attained knowledge about the language, the linguist next makes decisions about the division of words into morphemes and the contributions made by individual morphemes to the meaning of the word or of the sentence. IGT efficiently brings together and presents all of this information.

In the traditional four-line IGT format, morphemes appear on one line and glosses for those morphemes on the next. The gloss line includes both labels for grammatical morphemes (e.g. `PL` or `COM`) and translations of stems (e.g. *salir* or *ropa*). See the following example from Uspanteko:[4]

[3] Spanish: *Solo asi yo aprendi con él. No le hable en el idioma mio. Si no que en el idioma su papá le hablo. Y solo asi me fui acostumbrando. Solo asi ahora yo platico con ellos.*

[4] KEY: E1S=singular first person ergative, INC=incompletive, PART=particle, PREP=preposition, PRON=pronoun, NEG=negation,

(2) `Kita' tinch'ab'ej laj inyolj iin`

(3)
kita'	t-in-ch'abe-j	laj	in-yolj	iin	
NEG	INC-E1S-hablar-SC	PREP	E1S-idioma	yo	
PART	TAM-PERS-VT-SUF	PREP	PERS-S	PRON	

'No le hablo en mi idioma.'
('I don't speak to him in my language.')

Most commonly, IGT is presented in a four-tier format. The first tier (2) is the raw, unannotated text. The second (first line of (3)) is the same text with each word morphologically segmented. The third tier (second line of (3)), the gloss line, is a combination of Spanish translations of the Uspanteko stems and gloss tags representing the grammatical information encoded by affixes and stand-alone morphemes. The fourth tier (fourth line of (3)) is a translation in the target language of documentation.

Some interlinear texts include other project-defined tiers. OKMA uses a fifth tier (third line of (3)), described as the word-class line. This line is a mix of traditional POS tags, positional labels (e.g. suffix, prefix), and broader linguistic categories like `TAM` for tense-aspect-mood.

2.3 Cleaning up the OKMA annotations

The OKMA annotations were created using Shoebox,[5] a standard tool used by documentary linguists for lexicon management and IGT creation. To develop a corpus suitable for these studies, it was necessary to put considerable effort into normalizing the original OKMA source annotations. Varied levels of linguistic training of the original annotators led to many inconsistencies in the original annotations. Also, Shoebox (first developed in 1987) uses a custom, pre-XML whitespace delimited data format, making normalization especially challenging. Finally, not all of the texts are fully annotated. Almost half of the 67 texts are just transcriptions, several texts are translated but not further analyzed, and several others are only partially annotated at text level, clause level, word level, or morpheme level. It was thus necessary to identify complete texts for use in our experiments. Some missing labels in nearly-complete texts were filled in by the expert annotator.

A challenge for representing IGT in a machine-readable format is maintaining the links between the source text morphemes in the second tier and the morpheme-by-morpheme glosses in the third tier. The standard Shoebox output format, for example, enforces these links through management of the number of spaces between items in the output. To address this, we converted the cleaned annotations into IGT-XML (Palmer and Erk, 2007) with help from the Shoebox/Toolbox interfaces provided in the Natural Language Toolkit (Robinson et al., 2007). Automating the transformation from Shoebox format to IGT-XML's hierarchical format required cleaning up tier-to-tier alignment and checking segmentation in some cases where morphemes and glosses were misaligned, as in (5) below.[6]

(4) `Non li in yolow rk'il`

(5)
Non	li	in	yolow	r-k'il	
DEM	DEM	yo	platicar	AP	E3s.-SR
DEM	DEM	PRON	VI	SUF	PERS SREL

'Solo asi yo aprendi con él.'

Here, the number of elements in the morpheme tier (first line of (5)) does not match the number of elements in the gloss tier (second line of (5)). The problem is a misanalysis of `yolow`: it should be segmented `yol-ow` with the gloss `platicar-AP`. Automating this transformation has the advantage of identifying such inconsistencies and errors.

There also were many low-level issues that had to be handled, such as checking and enforcing consistency of tags. For example, the tag `E3s.` in the gloss tier of (5) is a typo; the correct tag is `E3S`. The annotation tool used in these studies does not allow such inconsistencies to occur.

2.4 Target labels

There are two main tasks in producing IGT: word segmentation (determination of stems and affixes) and glossing each segment. Stems and affixes each get a different type of gloss: the gloss of a stem is typically its translation whereas the gloss of an affix is a label indicating its grammatical role. The additional word-class line provides part-of-speech information for the stems, such as VT for *salir*.

Complete prediction of segmentation, gloss translations and labels is our ultimate goal for aiding IGT

S=sustantivo (noun), SC=category suffix, SUF=suffix, TAM=tense/aspect/mood, VT=transitive verb

[5] `http://www.sil.org/computing/shoebox/`

[6] KEY: AP=antipassive, DEM=demonstrative, E3S=singular third person ergative, PERS=person marking, SR/SREL=relational noun, VI=intransitive verb

creation with automation. Here, we study the potential for improving annotation efficiency for the more limited task of predicting the gloss label for each affix and the part-of-speech label for each stem. Thus, the experiments aim to produce a single label for each morpheme. We assume that words have been pre-segmented and we ignore the gloss translations.

The target representation in these studies is an additional tier which combines gloss labels for affixes and stand-alone morphemes with part-of-speech labels for stems. Example (6) repeats the clause in (4), adding this new combined tier. Stem labels are given in bold text, and affix labels in plain text.

(6) `Non li in yolow rk'il`

(7) `Non li in yol-ow r-k'il`
DEM DEM **PRON** **VI**-AP E3S-SR

'Solo asi yo aprendi con él.'

A simple procedure was used to create the new tier. For each morpheme, if a gloss label (such as `DEM` or `E3S`) appears on the gloss line (second line of (3)), we select that label. If what appears is a stem translation, we instead select the part-of-speech label from the next tier down (third line of (3)).

In the entire corpus, sixty-nine different labels appear in this combined tier. The following table shows the five most common part-of-speech labels (left) and the five most common gloss labels (right). The most common label, `S`, accounts for 11.3% of the tokens in the corpus.

S	noun	7167	**E3S**	sg.3p. ergative	3433
ADV	adverb	6646	**INC**	incompletive	2835
VT	trans. verb	5122	**COM**	completive	2586
VI	intrans. verb	3638	**PL**	plural	1905
PART	particle	3443	**SREL**	relational noun	1881

3 Integrated annotation and automation

The experimental framework described in this section is designed to model and evaluate real-time integration of human annotation, active learning strategies, and output from machine-learned classifiers. The task is annotation of morpheme-segmented texts from a language documentation project (sec. 2).

3.1 Tools and resources

Integrating automated support and human annotation in this context requires careful coordination of three components: 1) presenting examples to the annotator and storing the annotations, 2) training and evaluation of tagging models using data labeled by the annotator, and 3) selecting new examples for annotation. The processes are managed and coordinated using the OpenNLP IGT Editor.[7] The annotation component of the tool, and in particular the user interface, is built on the Interlinear Text Editor (Lowe et al., 2004).

For tagging we use a strong but simple standard classifier. There certainly are many other modeling strategies that could be used, for example a conditional random field (as in Settles and Craven (2008)), or a model that deals differently with POS labels and morpheme gloss labels. Nonetheless, a documentary linguistics project would be most likely to use a straightforward, off-the-shelf labeler, and our focus is on exploring different annotation approaches in a realistic documentation setting rather than building an optimal classifier. To that end, we use a standard maximum entropy classifier which predicts the label for a morpheme based on the morpheme itself plus a window of two morphemes before and after. Standard features used in part-of-speech taggers are extracted from the morpheme to help with predicting labels for previously unseen stems and morphemes.

3.2 Annotators and annotation procedures

A practical goal of these studies is to explore best practices for using automated support to create fully-annotated texts of the highest quality possible within fixed resource limits. For producing IGT, one of the most valuable resources is the time of a linguist with language-specific expertise. Documentary projects may also (or instead) have access to a trained linguist without prior experience in the language. We compare results from two annotators with different levels of exposure to the language. Both are trained linguists who specialize in language documentation and have extensive field experience.[8]

The first, henceforth referred to as the **expert annotator**, has worked extensively on Uspanteko, including writing a grammar of the language and

[7] `http://igt.sourceforge.net/`

[8] It should be noted that these are pilot studies. With just two annotators, the annotation comparisons are suggestive but not conclusive. Even so, this scenario accurately reflects the resource limitations encountered in documentation projects.

contributing to the publication of an Uspanteko-Spanish dictionary (Ángel Vicente Méndez, 2007). She is a native speaker of K'ichee', a closely-related Mayan language. The second annotator, the **non-expert annotator**, is a doctoral student in language documentation with no prior experience with Uspanteko and only limited previous knowledge of Mayan languages. Throughout the annotation process, the non-expert annotator relied heavily on the Uspanteko-Spanish dictionary. Both annotators are fluent speakers of Spanish, the target translation and glossing language for the OKMA texts.

In many annotation projects, labeling of training data is done with reference to a detailed annotation manual. In the language documentation context, a more usual situation is for the annotator(s) to work from a set of agreed-upon conventions but without strict annotation guidelines. This is not because documentary linguists lack motivation or discipline but simply because many aspects of the language are unknown and the analysis is constantly changing.

In the absence of explicit written annotation guidelines, we use an annotation training process for the annotators to learn the OKMA annotation conventions. Two seed sets of ten clauses each were selected to be used both for human annotation training and for initial classifier training. The first ten clauses of the first text in the training data were used to seed model training for the sequential selection cases (see 3.4). The second set of ten were randomly selected from the entire corpus and used to seed model training for both random and uncertainty sampling.

These twenty clauses were used to provide initial guidance to the annotators. With the aid of a list of possible labels and the grammatical categories they correspond to, each annotator was asked to label the seed clauses, and these labels were compared to the gold standard labels. Annotators were told which labels were correct and which were incorrect, and the process was repeated until all morphemes were correctly labeled. In some cases during this training phase, the correct label for a morpheme was supplied to the annotator after several incorrect guesses.

3.3 Suggesting labels

We consider two situations with respect to the contribution of the classifier: a **suggest** condition in which the labels predicted by the machine learner are shown to the annotator as she begins labeling a selected clause, and a **no-suggest** condition in which the annotator does not see the predicted labels.

In the suggest cases, the annotator is shown the label assigned the greatest likelihood by the tagger as well as a list of several highly-likely labels, ranked according to likelihood. To be included on this list, a label must be assigned a probability greater than half that of the most-likely label. In the no-suggest cases, the annotator has access to a list of the labels previously seen in the training data for a given morpheme, ranked in order of frequency of occurrence with the morpheme in question; this is similar to the input an annotator gets while glossing texts in Shoebox/Toolbox. Specifically, Shoebox/Toolbox presents previously seen glosses and labels for a given morpheme in alphabetic order.

3.4 Sample selection

We consider three methods of selecting examples for annotation–sequential (**seq**), random (**rand**), and uncertainty sampling (**al**)–and the performance of each method in both the **suggest** and the **no-suggest** setups. For uncertainty sampling, we measure uncertainty of a clause as the average entropy per morpheme (i.e., per labeling decision).

3.5 Measuring annotation cost

Not all examples take the same amount of effort to annotate. Even so, the bulk of the literature on active learning assumes some sort of unit cost to determine the effectiveness of different sample selection strategies. Examples of unit cost measurements include the number of documents in text classification, the number of sentences in part-of-speech tagging (Settles and Craven, 2008), or the number of constituents in parsing (Hwa, 2000). These measures are convenient for performing active learning simulations, but awareness has grown that they are not truly representative measures of the actual cost of annotation (Haertel et al., 2008a; Settles et al., 2008), with Ngai and Yarowsky (2000) being an early exception to the unit-cost approach. Also, Baldridge and Osborne (2004) use discriminants in parse selection, which are annotation decisions that they later showed correlate with timing information (Baldridge and Osborne, 2008).

The cost of annotation ultimately comes down to

money. Since annotator pay may be variable but will (under standard assumptions) be constant for a given annotator, the best approximation of likely cost savings is to measure the time taken to annotate under different levels of automated support. This is especially important in sample selection and its interaction with automated suggestions: active learning seeks to find more informative examples, and these will most likely involve more difficult decisions, decreasing annotation quality and/or increasing annotation time (Hachey et al., 2005). Thus, we measure cost in terms of the time taken by each annotator on each example. This allows us to measure the actual time taken to produce a given labeled data set, and thus compare the effectiveness of different levels of automated support plus their interaction with annotators of different levels of expertise.

Recent work shows that paying attention to predicted annotation cost in sample selection itself can increase the effectiveness of active learning (Settles et al., 2008; Haertel et al., 2008b). Though we have not explored cost-sensitive selection here, the scenario described here is an appropriate test ground for it: in fact, the results of our experiments, reported in the next section, provide strong evidence for a real natural language annotation task that active learning selection with cost-sensitivity is indeed sub-optimal.

4 Discussion

This section presents and discusses preliminary results from the ongoing annotation experiments. The Uspanteko corpus was split into training, development, and held-out test sets, roughly 50%, 25%, and 25%. Specifically, the training set of 21 texts contains 38802 words, the development set of 5 texts contains 16792 words, and the held-out test set, 6 texts, contains 18704 words. These are small datasets, but the size is realistic for computational work on endangered languages.

When measuring the performance of annotators, factors like fatigue, frustration, and especially the annotator's learning process must be considered. Annotators improve as they see more examples (especially the non-expert annotator). To minimize the impact of the annotator's learning process on the results, annotation is done in rounds. Each round consists of ten clauses from each of the six experimental

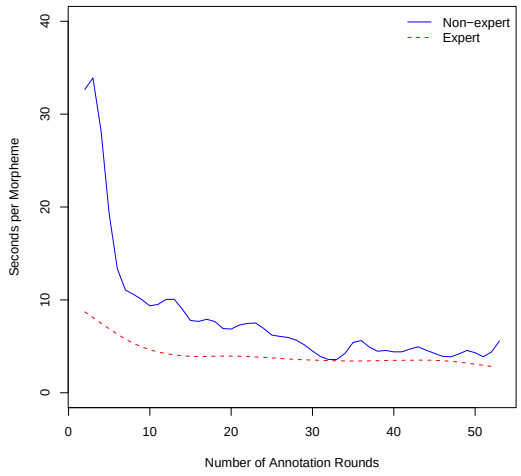

Figure 1: Average annotation time (in seconds per morpheme) over annotation rounds, averaged over all six conditions for each annotator.

cases for each annotator. The newly-labeled clauses are then added to the labeled training data, and a new tagging model is trained on the updated training set and evaluated on the development set. Both annotators have completed fifty-one rounds of annotation so far, labeling 510 clauses for each of the six experimental conditions. The average number of morphemes labeled is 3059 per case. Because the annotation experiments are ongoing, we discuss results in terms of the trends seen thus far.

4.1 Annotator speed

The expert annotator showed a small increase in speed after an initial familiarization period, and the non-expert showed a dramatic increase. Figure 1 plots the number of seconds taken per morpheme over the course of annotation, averaged over all six conditions for each annotator. The slowest, fastest, and mean rates, in seconds per morpheme, for the expert annotator were 12.60, 1.89, and 4.14, respectively. For the non-expert, they were 59.71, 1.90, and 8.03.

4.2 Accuracy of model on held-out data

Table 1 provides several measures of the current state of annotation in all 12 conditions after 51 rounds of annotation. The sixth column, labeled

Anno	Suggest	Select	Time (sec)	#Morphs	Model Accuracy	Total Accuracy of Annotation
NonExp	N	Seq	23739.79	3314	63.28	63.92
NonExp	N	Rand	22721.11	2911	**68.36**	68.69
NonExp	N	AL	23755.71	2911	68.26	67.84
NonExp	Y	Seq	21514.05	2887	66.55	66.89
NonExp	Y	Rand	22189.68	3002	**68.41**	68.73
NonExp	Y	AL	25731.57	2750	67.63	67.30
Exp	N	Seq	11862.39	3354	61.15	61.88
Exp	N	Rand	11665.10	3043	64.60	64.91
Exp	N	AL	13894.14	3379	**66.74**	66.47
Exp	Y	Seq	11758.74	2892	61.12	61.48
Exp	Y	Rand	11426.85	2979	60.13	60.57
Exp	Y	AL	16253.40	3296	**63.30**	63.15

Table 1: After 51 rounds of annotation: ModelAcc=accuracy on development set, TotalAnnoAcc=accuracy of fully-labeled corpus

ModelAcc, shows the accuracy of models on the development data. This represents a unit cost assumption at the clause level: measured this way, the results would suggest that the non-expert was best served by random selection, with no effect from machine suggestions. For the expert, they suggest active learning without suggestions is best, and that suggestions actually hurt effectiveness.

4.3 Accuracy of fully-labeled corpus

We are particularly concerned with the question of how to develop a fully-labeled corpus with the highest level of accuracy, given a finite set of resources. Thus, we combine the portion of the training set labeled by the human annotator with the results of tagging the *remainder* of the training set with the model trained on those annotations. The rightmost column of Table 1, labeled **Total Accuracy of Annotation**, shows the accuracy of the fully labeled training set (part human, part machine labels) after 51 rounds. These accuracies parallel the model accuracies: random selection is best for the non-expert annotator, and uncertainty selection is best for the expert.

Since this tagging task involves labeling morphemes, a clause cost assumption is not ideal—e.g., active learning tends to select longer clauses and thereby obtains more labels. To reflect this, a subclause cost can help: here we use the number of morphemes annotated. The column labeled **Tokens** in Table 2 shows the total accuracy achieved in each condition when human annotation ceases at 2750 morphemes. The figure in parentheses is the cumulative annotation time at the morpheme cut-off point. Here, the non-expert does best: he took great care with the annotations and was clearly not tempted to accept erroneous suggestions from the machine labeler. In contrast, the expert does seem to have accepted many bad machine suggestions.

Anno	Suggest	Select	Time (11427 sec)	Tokens (time) (2750 morphs)
NonExp	N	Seq	55.01	59.80 (21678 secs)
NonExp	N	Rand	59.95	**68.68** (22069 secs)
NonExp	N	AL	59.86	67.70 (22879 secs)
NonExp	Y	Seq	60.27	66.79 (21053 secs)
NonExp	Y	Rand	62.96	68.38 (21194 secs)
NonExp	Y	AL	59.18	67.30 (25732 secs)
Exp	N	Seq	61.21	59.18 (10110 secs)
Exp	N	Rand	64.92	64.42 (10683 secs)
Exp	N	AL	**65.72**	65.74 (11826 secs)
Exp	Y	Seq	61.47	61.47 (11436 secs)
Exp	Y	Rand	60.57	61.16 (10934 secs)
Exp	Y	AL	61.54	62.87 (13957 secs)

Table 2: For given cost, accuracy of fully-labeled corpus.

Morpheme unit cost is more fine-grained than clause-level cost, but it hides the fact that the expert annotator needed far less time to produce a corpus of higher overall labeled quality than the non-expert. This can be seen in the **Time** column of Table 2, which gives the total annotation accuracy when 11427 seconds are alloted for human labeling. The expert annotator achieved the highest accuracy for total labeling of the training set using active learning without machine label suggestions. Active learning helps the non-expert as well, but his best condition is random selection with machine labels.

4.4 Annotator accuracy by round

Active learning clearly selects harder examples that hurt the non-expert's performance. To see this clearly, we measured the accuracy of the annotators' labels for each round of each experimental setup,

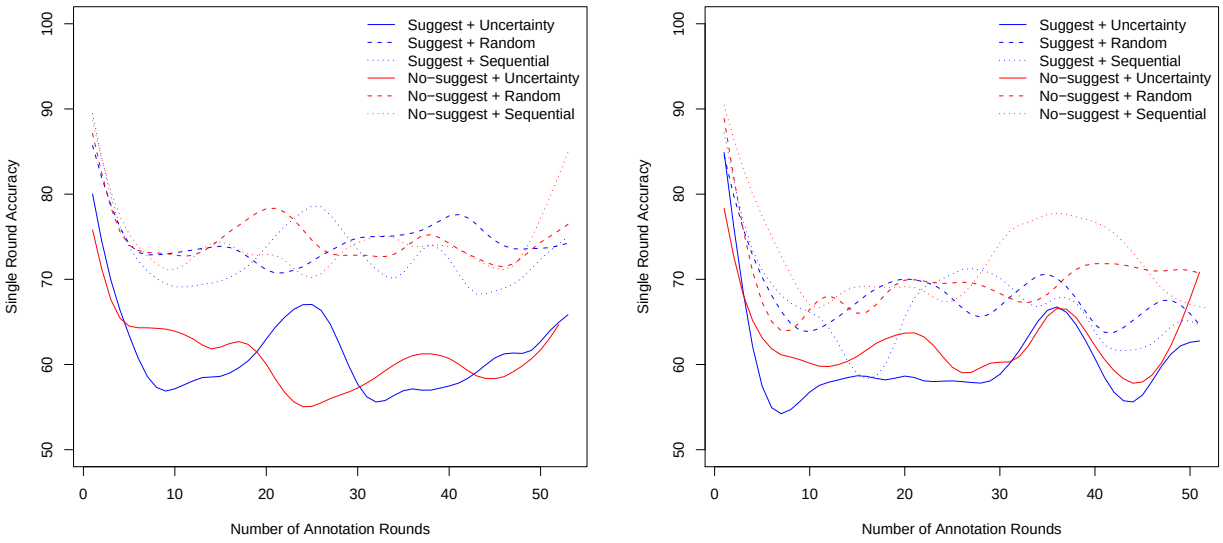

Figure 2: Single round accuracy per round for each experiment type by: (a) non-expert annotator, (b) expert annotator

given in Fig. 2. It is not clear at this stage whether the tag suggestions by the machine labeler are helpful to human annotation. It is useful to compare the cases where the machine learner is not involved in example selection (i.e. random and sequential) to uncertainty sampling, which does involve the machine learner. One thing that is apparent is that when active learning is used to select samples for annotation, both the expert and non-expert annotator have a harder time providing correct tags. A point of contrast between the expert and non-expert is that the non-expert generally outperforms the expert on label accuracy in the non-active learning scenarios. The non-expert was very careful with his labeling decisions, but also much slower than the expert. In the end, speedier annotation rates allowed the expert annotator to achieve higher accuracies in less time.

5 Conclusion

We have described a set of ongoing pilot experiments designed to test the utility of machine labeling and active learning in the context of documentary linguistics. The production of IGT is a realistic annotation scenario which desperately needs labeling efficiency improvements. Our preliminary results suggest that both machine labeling and active learning can increase the effectiveness of annotators, but they interact quite strongly with the expertise of the annotators. In particular, though active learning works well with the expert annotator, for a non-expert annotator it seems that random selection is a better choice. However, we stress that our annotation experiments are ongoing. Active learning is often less effective early in the learning curve, especially when automated label suggestions are provided, because the model is not yet accurate enough to select truly useful examples, nor to suggest labels for them reliably (Baldridge and Osborne, 2004). Thus, we expect automation via uncertainty sampling and/or suggestion may gather momentum and outpace random selection and/or no suggestions by wider margins as annotation continues.

Acknowledgments

This work is funded by NSF grant BCS 06651988 “Reducing Annotation Effort in the Documentation of Languages using Machine Learning and Active Learning.” Thanks to Katrin Erk, Nora England, Michel Jacobson, and Tony Woodbury; and to annotators Telma Kaan Pixabaj and Eric Campbell. Finally, thanks to the anonymous reviewers for valuable feedback.

References

Miguel Ángel Vicente Méndez. 2007. *Diccionario bilingüe Uspanteko-Español. Cholaj Tzijb'al li Uspanteko*. Okma y Cholsamaj, Guatemala.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of Empirical Approaches to Natural Language Processing (EMNLP)*.

Jason Baldridge and Miles Osborne. 2008. Active learning and logarithmic opinion pools for HPSG parse selection. *Natural Language Engineering*, 14(2):199–222.

David Crystal. 2000. *Language Death*. Cambridge University Press, Cambridge.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, Ann Arbor, MI.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and McClanahan Peter. 2008a. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 65–68, Columbus, Ohio, June. Association for Computational Linguistics.

Robbie A. Haertel, Kevin D. Seppi, Eric K. Ringger, and James L. Carroll. 2008b. Return on investment for active learning. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*. ACL Press.

Rebecca Hwa. 2000. Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT Conference on EMNLP and VLC*, pages 45–52, Hong Kong, China, October.

William Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of IJCNLP-2008*, Hyderabad, India.

John Lowe, Michel Jacobson, and Boyd Michailovsky. 2004. Interlinear text editor demonstration and projet archivage progress report. In *4th EMELD workshop on Linguistic Databases and Best Practice*, Detroit, MI.

Grace Ngai and David Yarowsky. 2000. Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 117–125, Hong Kong.

Alexis Palmer and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop (LAW-07), ACL07*, Prague.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, María Vicente Méndez, and Oswaldo Ajcot Damián. 2007. Text collections in Four Mayan Languages. Archived in The Archive of the Indigenous Languages of Latin America.

Michael Richards. 2003. *Atlas lingüístico de Guatemala*. Servipresna, S.A., Guatemala.

Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with Toolbox and the Natural Language Toolki t. *Language Documentation and Conservation*, 1:44–57.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii, October. Association for Computational Linguistics.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1069–1078. ACL Press.

Fei Xia and William Lewis. 2007. Multilingual structural projection across interlinear text. In *Proceedings of HLT/NAACL 2007*, Rochester, NY.

Fei Xia and William Lewis. 2008. Repurposing theoretical linguistic data for tool development antd search. In *Proceedings of IJCNLP-2008*, Hyderabad, India.