Tag Dictionaries Accelerate Manual Annotation

Marc Carmen*, Paul Felt†, Robbie Haertel†, Deryle Lonsdale*, Peter McClanahan†, Owen Merkling†, Eric Ringger†, Kevin Seppi†

*Department of Linguistics and †Department of Computer Science
Brigham Young University
Provo, Utah 84602 USA
marc.carmen@gmail.com, pablofelt@gmail.com, robbie_haertel@byu.edu, lonz@byu.edu,
petermcclanahan@gmail.com, omerkling@gmail.com, ringger@cs.byu.edu, kseppi@byu.edu
http://nlp.cs.byu.edu/

Abstract

Expert human input can contribute in various ways to facilitate automatic annotation of natural language text. For example, a part-of-speech tagger can be trained on labeled input provided offline by experts. In addition, expert input can be solicited by way of active learning to make the most of annotator expertise. However, hiring individuals to perform manual annotation is costly both in terms of money and time. This paper reports on a user study that was performed to determine the degree of effect that a part-of-speech dictionary has on a group of subjects performing the annotation task. The user study was conducted using a modular, web-based interface created specifically for text annotation tasks. The user study found that for both native and non-native English speakers a dictionary with greater than 60% coverage was effective at reducing annotation time and increasing annotator accuracy. On the basis of this study, we predict that using a part-of-speech tag dictionary with coverage greater than 60% can reduce the cost of annotation in terms of both time and money.

1. Introduction

For many computational and corpus linguistics tasks an annotated corpus is required. Unfortunately manual annotation is often cost-prohibitive; consequently, only a modest portion of large corpora can feasibly be annotated without additional assistance. Various tools have been created to simplify and accelerate the manual annotation process while maintaining annotation quality.

Many annotation tasks involve per-token tags. Perhaps the most common of such tasks is part-of-speech (POS) tagging, but the entity names in Named Entity Recognition and the constituent labels in parse trees, to name a few, can also be seen as a type of tag.

Tag dictionaries are one method of accelerating manual annotation for such tasks. Tag dictionaries contain a list of possible tags for a particular instance requiring annotation. For instance, in POS tagging, for a given token, a human annotator can be prompted with a list of all possible tags, or the annotator can be given an appropriate subset of tags to choose from. Typically the subset will be selected based on tags previously assigned to the type for the token in question. If the subset of tags in this tag dictionary is substantially smaller than the full list *and* contains the correct tag, we might expect the tag dictionary to reduce the amount of time it takes to find and select the correct answer. Having fewer options may also improve the annotator's ability to select the correct one. On the other hand, if the tag dictionary does not contain the correct tag, it will presumably take even more effort to discover this, take the steps necessary to show a complete list of tags, and select the answer from that list instead. Likewise, if the tag dictionary contains most of the tags, it is not likely to save much time and may even require additional time.

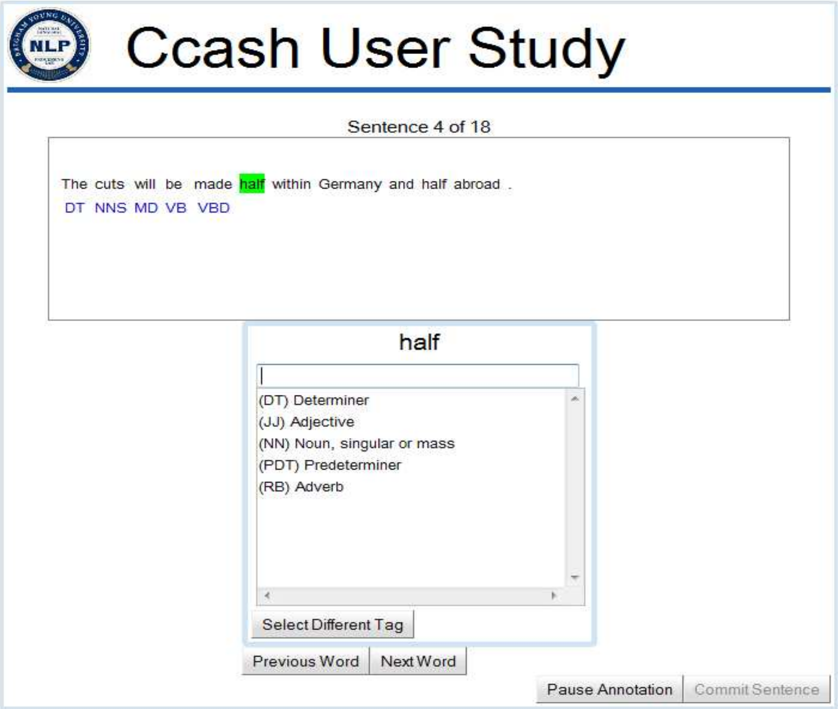
We report on a user study in which we employ a part-of-speech (POS) tag dictionary for an English POS annotation task, and we quantify the degree to which a tag dictionary can accelerate and improve the annotation process in terms of time and accuracy.

This paper is organized as follows: we begin with a review of related work in Section 2 which looks at a variety of tools that have been created for POS tagging. Section 3 discusses the overall design of the project including the user interface and the details of the data being used and its preparation. Section 4 provides the results of the user study that was performed. The final section provides an overview of the results as well as some ideas regarding possible future directions.

2. Related Work

Many tools have been created to facilitate the annotation process and thereby to reduce the cost of annotation. For example, Knowtator (a plug-in to Protégé, an ontology editor) allows for easy creation of complex annotation schemas that include features like "constrained relationships between the [annotation] types" (Ogren, 2006); Knowtator's annotation schema can be applied to an annotation task, which means that no coding is required for new tasks. WordFreak is an extensible annotation system that allows for easy integration of additional components and new annotation tasks (Morton & LaCivita, 2003). In addition, it provides access to several automatic tools including sentence detectors, part-of-speech taggers, and parsers. Rayson, et al. (2006) proposed a peer-to-peer framework to distribute the work of part-of-speech tagging and shallow parsing and to tie into existing tools such as the CLAWS tagger.

These tools and many others attempt to lower the cost of corpus annotation, although quantifying their actual reduction in cost requires additional analysis as well as



Ccash User Study

Sentence 4 of 18

The cuts will be made half within Germany and half abroad .
DT NNS MD VB VBD

half

(DT) Determiner
(JJ) Adjective
(NN) Noun, singular or mass
(PDT) Predeterminer
(RB) Adverb

Figure 1. Interface used for POS tagging. The first five words have been tagged and the selected tag is shown below. This is an example where a dictionary is suggesting possible tags.

user studies. Some user studies have reported improvements in either time or accuracy during the annotation process. Ringger, et al. (2008) conducted a user study which allowed the authors to derive a model of the time (and therefore cost) required for English POS annotation with the aid of active learning. The authors presented linear cost models for both word-at-a-time and sentence-at-a-time active learning-based annotation. Palmer, Moon, and Baldridge (2009) conducted a user study involving automatic pre-annotation and active learning with both an expert and non-expert annotating the Uspanteko language with inter-linear glosses. They found that machine labeling and active learning can increase the accuracy of human annotators but that the degree to which they increase the accuracy is related to the experience of the annotator. Culotta et al. (2005) measure the cost of annotation in terms of the number of required user actions to fix an annotation in a user interface incorporating automatic pre-annotation. To our knowledge, no study has been conducted on the effectiveness of tag dictionaries in reducing annotation costs.

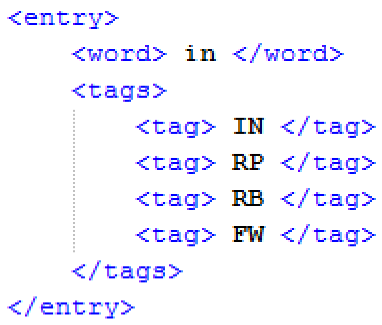
3. Experimental Design

The purpose of the study documented in the present work is to quantify the effectiveness of tag dictionaries on improving accuracy and reducing annotation time. In order to do so, we employ a user study which measures annotation time and accuracy in a part-of-speech tagging task both with and without tag dictionaries. Since the results are highly influenced by the user interface, the data, and the participants, we detail each of these in turn.

3.1 User Interface

The user study was built using an annotation framework of our own design called CCASH (Cost Conscious Annotation Supervised by Humans; Felt et al., 2010). As documented elsewhere in this proceedings, CCASH is a web-based annotation tool that facilitates corpus annotation with cost reduction in mind. Built using the Google Web Toolkit (GWT), Hibernate, and MySQL, CCASH provides annotators with a browser-based annotation client. Depending on its configuration, the tool allows for straight-forward manual annotation, annotation using suggestions from a tag dictionary, automatic pre-annotation using machine-learned models, and/or sample selection using an active learning algorithm. To maximize the efficiency of annotators, both keyboard and mouse interaction options are available, and the interface was designed for streamlined interaction involving a minimal number of mouse clicks.

In the study reported here, CCASH was used to measure the ability of a POS tag dictionary to improve the annotation process with regard to both the speed and accuracy of annotation. For every token, the CCASH interface presents a list of tag options contained in the tag dictionary (e.g., see Figure 1). For those tokens not covered by the current tag dictionary, all possible tags are listed. If a dictionary contains a sufficiently complete tag inventory for a given token, the limited options for that token make the choice potentially easier for the annotator. If a dictionary entry is missing from the list, the annotator



```
<entry>
    <word> in </word>
    <tags>
        <tag> IN </tag>
        <tag> RP </tag>
        <tag> RB </tag>
        <tag> FW </tag>
    </tags>
</entry>
```

Figure 2. Example tag dictionary entry for the word 'in'

can add that option to the dictionary. The trade-off is that although a dictionary has the potential to accelerate annotation, an incomplete dictionary may require additional effort to augment. This is particularly the case in our user interface since the user must click a button (labeled “Select Different Tag” in Figure 1) and choose from a second list when the option was not in the original. We also note the potential for tag dictionaries to affect accuracy: a complacent annotator may choose an option simply because it is the best in the list rather than considering the full range of options. On the other hand, the limited list may help the annotator reduce the range of options and choose the correct tag.

3.2 Data

Our data comes from the Wall Street Journal section of the Penn Treebank (Marcus et al, 1994). We present sentences from the Treebank to the users, but we also use the Treebank to construct our tag dictionaries. The scenario for the study involves a human annotator tagging text and simulating the addition of entries to the tag dictionary as needed (in the study, these additions do not persist to future word tokens). In a real annotation task, the content of a tag dictionary will lie on a spectrum from an empty tag dictionary, a partial tag dictionary, up to a complete tag dictionary. We assume that tag dictionaries consist of all tags previously assigned to the current word type. Therefore, the tag dictionary for a given type evolves during the annotation process. This study snapshots various stages of the annotation process over a large corpus in order to assess at what point of coverage the dictionary can actually help.

The sentences in this study (18 in all) were randomly selected from the Penn Treebank such that sentences were of three lengths: short (12 tokens), medium (23 tokens), and long (36 tokens); 6 sentences were selected per length bucket. For each sentence, POS tag dictionaries were constructed such that they contained tags for approximately 20%, 40%, 60%, 80%, or 100% of the tokens in order to simulate the coverage of the dictionaries at various stages of annotation. The dictionary coverage level is determined using the following simple formula:

$$coverage = \frac{\#\ of\ tokens\ with\ at\ least\ one\ entry}{\#\ of\ tokens\ in\ the\ sentence}$$

Note that 100% coverage level does not imply a complete tag dictionary, simply that all tokens have at least one entry. Figure 2 shows an example of an entry for the word “in” in a tag dictionary with 80% coverage. The tag entries in this example, as in the rest of the study, are listed in the order in which they first occur in the shuffled corpus.

Dictionaries were created automatically offline in order to achieve a desired coverage level. Creating tag dictionaries involved mimicking a realistic annotation process in which tag dictionaries grow over time when new entries are added. An additional advantage of this approach besides creating realistic dictionaries is that more common words (including function words) have at least one tag entry in each of the dictionaries. Each dictionary was created by first removing the tutorial and user study sentences from the Penn Treebank dataset. Then the dataset was randomly shuffled and split into training and held-out sets of equal size. A base dictionary for each coverage level was created by adding each word of every sentence in the training set and its annotation until the average coverage for a dictionary on the entire held-out dataset reached the desired coverage level (20%, 40%, 60%, or 80%). Once a base dictionary was created, a dictionary for each of the 18 user study sentences and each coverage level was created as follows: starting with the base dictionary for a given coverage level, tagged words from sentences in the training set were either added or removed in order, starting from the last sentence used to create the base dictionary. The process was stopped when the difference between the actual coverage level for that dictionary and the target coverage level was a minimum. The method produced tag dictionaries whose actual coverage levels are within 2.12% (absolute) of the specified coverage level on average.

Participants were presented the same sentences in the same order, this order having been pre-established randomly. However, the coverage level of the dictionary was randomized for each sentence presented to the participant, under the constraint that a given user be assigned a unique coverage level for each of the 6 sentences in every length bucket. This method ensured that each sentence was annotated at a maximum number of distinct coverage levels by different participants. Using this method, it was possible that a user would be presented with a sentence with a high coverage dictionary immediately followed by a sentence with lower coverage. Unfortunately, this process does not properly mimic the evolution of tag dictionaries in a real annotation task, but it does allow us to discriminate the effects of coverage level from sentence order on time and accuracy.

Before beginning the study, participants were presented with a tutorial consisting of four randomly selected sentences. This helped mitigate the potential effects of the human learning curve and, in addition, it helped familiarize the participants with the user interface and the Penn Treebank tag set. Each participant was provided with a one-sheet table of the tags including a list of examples for each tag. During the tutorial each participant was asked to annotate the same set of sentences and was provided the same tag dictionary coverage for each tutorial sentence. After annotating each sentence and submitting their annotations, the participants were shown the correct annotations and given visual feedback about which annotations were correct and which were not. The participants were not allowed to move past a tutorial sentence until they had correctly annotated at least all but one of the tokens in the sentence.

			Time					Accuracy				
Length	**Coverage**	**Num**	**Min**	**Mean**	**Max**	**p-val**	**Perm**	**Min**	**Mean**	**Max**	**p-val**	**Perm**
12	Full Dict	31	54	106	174	0.50	1.00	0.50	0.80	1.00	0.50	1.00
	20	31	48	136	238	0.79	0.41	0.58	0.81	1.00	0.43	0.87
	40	33	39	94	204	0.35	0.71	0.50	0.83	1.00	0.21	0.40
	60	29	40	100	139	**0.01**	**0.02**	0.67	0.83	1.00	0.18	0.37
	80	32	30	94	204	0.24	0.49	0.75	0.86	1.00	**0.00**	**0.01**
	100	31	26	85	133	**0.00**	**0.01**	0.75	0.86	1.00	**0.01**	**0.01**
23	Full Dict	27	64	258	264	0.50	1.00	0.70	0.87	1.00	0.50	1.00
	20	31	88	191	309	0.86	0.31	0.70	0.86	0.96	0.76	0.50
	40	29	88	191	253	0.22	0.44	0.74	0.88	1.00	0.30	0.62
	60	30	66	160	257	**0.07**	0.17	0.83	0.87	0.96	**0.08**	0.18
	80	30	54	130	225	**0.00**	**0.00**	0.83	0.89	0.96	**0.01**	**0.03**
	100	31	52	121	202	**0.00**	**0.00**	0.83	0.90	1.00	**0.06**	0.13
36	Full Dict	33	121	265	533	0.50	1.00	0.75	0.88	1.00	0.50	1.00
	20	32	113	248	465	0.15	0.32	0.72	0.87	0.97	0.71	0.57
	40	32	93	282	577	0.32	0.65	0.75	0.90	1.00	0.16	0.33
	60	30	82	219	353	**0.00**	**0.00**	0.81	0.92	0.97	**0.00**	**0.00**
	80	28	85	204	310	**0.00**	**0.00**	0.81	0.93	1.00	**0.00**	**0.00**
	100	31	90	191	318	**0.00**	**0.00**	0.78	0.93	1.00	**0.00**	**0.00**

Figure 3. Results of t-test and permutation test. The “Num” column indicates the number of data points available for the condition. “Perm” is analagous to p-val, but for the permutation test. Significant (at confidence level 90% or higher) results are highlighted.

Thirty-three subjects participated in this user study. Subjects were first-year linguistics graduate students in a required syntax and morphology course. Responses to a questionnaire at the end of the study indicate that twenty-three of the participants are native English speakers, and over 50% of the students had taken one or fewer previous courses that cover POS tagging. In addition, when asked about their tagging proficiency, over 50% of the participants rated themselves with a 1 (lowest proficiency) or 2 out of 5 (highest). They were given an assignment by the instructor and were told that credit would be given based only on completion of the study and whether or not the results indicated that the participant had taken the study seriously: participants were told that both accuracy and time were important for the study.

4. Results

On average the participants performed the annotation task with 88.73% accuracy. The lowest accuracy score was 80.52%, and the highest accuracy score was 93.90% for the study. The non-native English speakers scored an average of 88.02% compared to the native speakers’ 88.96%. The participants required from 22.76 minutes to 118.43 minutes to complete the study, with an average of 42.63 minutes. On average, the non-native speakers took approximately 20 minutes longer than the native speakers to complete the study.

The most important results from this study concern the sentence-level statistics for each length of sentence and coverage level. Our study was designed under the hypothesis that tag dictionaries can accelerate annotation and make it more accurate compared to always presenting the user with all tags. Consequently, the null hypothesis is that having no dictionary (i.e., presenting the user with every tag) has the same effect on time and accuracy as having a tag dictionary. The time and accuracy for each sentence is analyzed using a *t*-test and permutation test (Menke & Martinez, 2004).

As a general trend, we see that the minimum and mean accuracy improve with increasing coverage level. Similarly, the minimum and mean time decrease with increasing coverage level. The accuracy also tends to be higher for longer sentences at the same coverage level. Though not with statistical significance, dictionaries with coverage levels of 20% tend to be worse than using the full dictionary. Low coverage dictionaries tend to have fewer entries and are therefore less likely to contain the correct tag. We suspect that the users trust these suggestions enough that they occasionally prefer the suggestions in the tag dictionary even though the correct answer is not present.

A similar trend exists for time, namely, that for dictionaries of 20% and 40% coverage level, the minimum, mean, and maximum times are greater than when using the full dictionary and short and medium length sentences. As previously noted, these dictionaries are less likely to contain the correct tag, and choosing a tag that is missing from the dictionary takes additional time. Although tag dictionaries are used less frequency in these scenarios (approximately 20% and 40% of the time), and despite the fact that the accuracy numbers suggest that the users occasionally avoided adding new entries, the additional time required to add entries may outweigh the benefits of having a tag dictionaries when coverage is low.

On the other hand, as the level of dictionary coverage increased there was a significant improvement in both time and accuracy. For each sentence length, statistically significant improvement occurred when dictionary coverage was at or above 60% with a confidence level of 80% or higher; however, most of the results were achieved with a confidence level of 95% or higher. Not surprisingly, a dictionary with 100% coverage was nearly always optimal showing improvement with a confidence level of 99% for most sentence lengths.

Finally, we note that improvements appear to be greater for longer sentences. This may be attributable to the fact that it is easier to build dictionaries close to the desired level of coverage for longer sentences. Another explanation for the differences in time may be that the overhead of tagging a sentence is a lower percentage of the total time for longer sentences.

5. Conclusions

Using a tag dictionary during the annotation process can speed up human annotation while still maintaining accuracy, if the dictionary has significant coverage (in our study, 60%). We plan to conduct additional user studies to test the validity of these and related ideas involving pre-annotation for other languages, including more morphologically complex languages. We are currently configuring CCASH for a study involving the morphological annotation of Syriac.

This user study and the work related to it has stemmed from previous work by the authors which utilizes active learning to reduce the cost of text annotation. In the future, the authors hope to introduce the active learning component and perform the same task with a human oracle and a dictionary.

Finally, part-of-speech annotation is just one form of text annotation. The authors have also worked on other components that can be utilized in the CCASH system. For example, one component developed would allow for annotators to perform named entity annotation.

6. References

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. *The Twentieth National Conference on Artificial Intelligence (AAAI),* (pp. 746--751). Pittsburgh, PA.

Felt, P., Merkling, O., Carmen, M., Ringger, E., Lemmon, W., Seppi, K., et al. (2010). CCASH: A Web Application Framework for Efficient Distributed Language Resource Development. *Proceedings of LREC 2010,* (p. this proceedings). Valetta, Malta.

Haertel, R. A., Seppi, K. D., Ringger, E. K., & Carroll, J. L. (2008a). Return on Investment for Active Learning. *Proceedings of NIPS Workshop on Cost Sensitive Learning.* Whistler, British Columbia, Canada.

Haertel, R., Ringger, E., Seppi, K., Carroll, J., & McClanahan, P. (2008b). Assessing the Costs of Sampling Methods in Active Learning for Annotation. *Proceedings of ACL-08: HLT, Short Papers* (pp. 65-68). Columbus: Association for Computational Linguistics.

Menke, J., & Martinez, T. R. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. *Proceedings of 2004 IEEE International Joint Conference on Neural Networks,* (pp. 1331-1335).

Morton, T., & LaCivita, J. (2003). WordFreak: an open tool for linguistic annotation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4,* (pp. 17-18). Edmonton, Alberta, Canada.

Ogren, P. V. (2006). Knowtator: a Protégé plug-in for annotated corpus construction. *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations,* (pp. 273-275). New York.

Palmer, A., Moon, T., & Baldridge, J. (2009). Evaluating automation strategies in language documentation. In E. Ringger, R. Haertel, & K. Tomanek (Ed.), *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing,* (pp. 36-44). Boulder, Colorado.

Rayson, P., Walkerdine, J., Fletcher, W. H., & Kilgariff, A. (2006). Annotated web as corpus. *Proceedings of the 2nd International Workshop on Web as Corpus.* Sydney, Australia.

Ringger, E., Carmen, M., Haertel, R., Seppi, K., Lonsdale, D., McClanahan, P., Carroll, J., Ellison, N. (2008). Assessing the costs of machine-assisted corpus annotation through a user study. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).* European Language Resources Association (ELRA).

Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., Seppi, K., Lonsdale, D. (2007). Active Learning for Part-of-Speech Tagging:Accelerating Corpus Annotation. *Proceedings of the 2007 ACL Linguistic Annotation Workshop,* (pp. 101-108). Prague, Czech Republic.