Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff

Matthias Sperber[1], Mirjam Simantzik[2], Graham Neubig[3], Satoshi Nakamura[3], Alex Waibel[1]
[1] arlsruhe nstitute of Technolog , nstitute for Anthropomatics, erman
[2] obile Technologies mb , erman
[3] ara nstitute of Science and Technolog , A C Laborator , apan
matthias.sperber@kit.edu, mirjam.simantzik@jibbigo.com, neubig@is.naist.jp
s-nakamura@is.naist.jp, waibel@kit.edu

Abstract

n this paper, e stud the problem of manuall correcting automatic annotations of natural language in as ef cient a manner as possible. e introduce a method for automaticall segmenting a corpus into chun s such that man uncertain labels are grouped into the same chun , hile human supervision can be omitted altogether for other segments. A tradeoff must be found for segment si es. Choosing short segments allo s us to reduce the number of highl con dent labels that are supervised b the annotator, hich is useful because these labels are often alread correct and supervising correct labels is a aste of effort. n contrast, long segments reduce the cognitive effort due to conte t s itches. ur method helps nd the segmentation that optimi es supervision ef cienc b de ning user models to predict the cost and utilit of supervising each segment and solving a constrained optimi ation problem balancing these contradictor ob ectives. A user stud demonstrates noticeable gains over pre-segmented, con dence-ordered baselines on t o natural language processing tas s: speech transcription and ord segmentation.

1 Introduction

an natural language processing (LP) tas s re uire human supervision to be useful in practice, be it to collect suitable training material or to meet some desired output ualit . iven the high cost of human intervention, ho to minimi e the supervision effort is an important research problem. Previous or s in areas such as active learning, post editing, and interactive pattern recognition have investigated this uestion ith notable success (Settles, 2008; Specia, 2011; on ale -Rubio et al., 2010).

(a) t as a bright cold **(they)** in **(apron)**, and **(a)** cloc s ere stri ing thirteen.

(b) t as a bright cold **(they)** in **(apron)**, and **(a)** cloc s ere stri ing thirteen.

(c) t as a bright cold **(they)** in **(apron)**, and **(a)** cloc s ere stri ing thirteen.

Figure 1: Three automatic transcripts of the sentence t as a bright cold da in April, and the cloc s ere stri ing thirteen , ith recognition errors in parentheses. The underlined parts are to be corrected b a human for (a) sentences, (b) ords, or (c) the proposed segmentation.

The most common frame or for ef cient annotation in the LP conte t consists of training an LP s stem on a small amount of baseline data, and then running the s stem on unannotated data to estimate con dence scores of the s stem s predictions (Settles, 2008). Sentences ith the lo est con dence are then used as the data to be annotated (Figure 1 (a)). o ever, it has been noted that hen the LP s stem in uestion alread has relativel high accurac , annotating entire sentences can be asteful, as most ords ill alread be correct (Tomane and ahn, 2009; eubig et al., 2011). n these cases, it is possible to achieve much higher bene t per annotated ord b annotating sub-sentential units (Figure 1 (b)).

o ever, as Settles et al. (2008) point out, simpl ma imi ing the bene t per annotated instance is not enough, as the real supervision effort varies

Transactions of the Association for Computational Linguistics, 2 (2014) 169–180. Action Editor: Eric Fosler-Lussier.
Submitted 11/2013; Revised 2/2014; Published 4/2014.

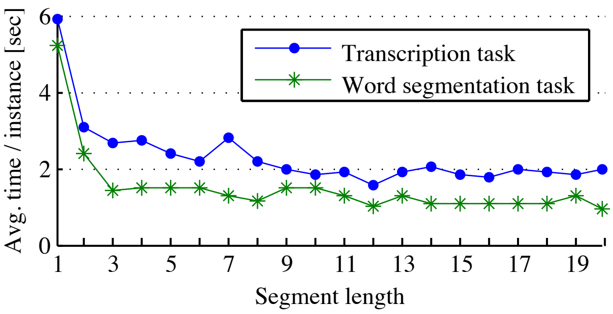

Figure 2: Average annotation time per instance, plotted over different segment lengths. For both tas s, the effort clearl increases for short segments.

greatl across instances. This is particularl important in the conte t of choosing segments to annotate, as human annotators heavil rel on semantics and conte t information to process language, and intuitivel , a consecutive se uence of ords can be supervised faster and more accuratel than the same number of ords spread out over several locations in a te t. This intuition can also be seen in our empirical data in Figure 2, hich sho s that for the speech transcription and ord segmentation tas s described later in Section 5, short segments had a longer annotation time per ord. ased on this fact, e argue it ould be desirable to present the annotator ith a segmentation of the data into easil supervisable chun s that are both large enough to reduce the number of conte t s itches, and small enough to prevent unnecessar annotation (Figure 1 (c)).

n this paper, e introduce a ne strateg for natural language supervision tas s that attempts to optimi e supervision ef cienc b choosing an appropriate segmentation. t relies on a user model that, given a speci c segment, predicts the cost and the utilit of supervising that segment. iven this user model, the goal is to nd a segmentation that mini mi es the total predicted cost hile ma imi ing the utilit . e balance these t o criteria b de ning a constrained optimi ation problem in hich one criterion is the optimi ation ob ective, hile the other criterion is used as a constraint. oing so allo s specif ing practical optimi ation goals such as remove as man errors as possible given a limited time budget, or annotate data to obtain some re uired classi er accurac in as little time as possible.

Solving this optimi ation tas is computationall dif cult, an P-hard problem. evertheless, e demonstrate that b ma ing realistic assumptions about the segment length, an optimal solution can be found using an integer linear programming formulation for mid-si ed corpora, as are common for supervised annotation tas s. For larger corpora, e provide simple heuristics to obtain an appro imate solution in a reasonable amount of time.

E periments over t o e ample scenarios demonstrate the usefulness of our method: Post editing for speech transcription, and active learning for apanese ord segmentation. ur model predicts noticeable ef cienc gains, hich are con rmed in e periments ith human annotators.

2 Problem Definition

The goal of our method is to nd a segmentation over a corpus of ord to ens w_1^N that optimi es supervision ef cienc according to some predictive user model. The user model is denoted as a set of functions $u_{l,k}(w_a^b)$ that evaluate an possible subse uence w_a^b of to ens in the corpus according to criteria $l \in L$, and supervision modes $k \in K$.

Let us illustrate this ith an e ample. Sperber et al. (2013) de ned a frame or for speech transcription in hich an initial, erroneous transcript is created using automatic speech recognition (ASR), and an annotator corrects the transcript either b correcting the ords b e board, b respea ing the content, or b leaving the ords as is. n this case, e could de ne $K=\{$T PE, RESPEA , S P$\}$, each constant representing one of these three supervision modes. ur method ill automaticall determine the appropriate supervision mode for each segment.

The user model in this e ample might evaluate ever segment according to t o criteria L, a cost criterion (in terms of supervision time) and a utilit criterion (in terms of number of removed errors), hen using each mode. ntuitivel , respea ing should be assigned both lo er cost (because spea ing is faster than t ping), but also lo er utilit than t ping on a e board (because respea ing recognition errors can occur). The S P mode denotes the special, unsupervised mode that al a s returns 0 cost and 0 utilit .

ther possible supervision modes include multiple input modalities (Suhm et al., 2001), several human annotators ith different e pertise and cost

(onme and Carbonell, 2008), and correction vs. translation from scratch in machine translation (Specia, 2011). Similarl , cost could instead be e pressed in monetar terms, or the utilit function could predict the improvement of a classi er hen the resulting annotation is not intended for direct human consumption, but as training data for a classi er in an active learning frame or .

3 Optimization Framework

iven this setting, e are interested in simultaneousl nding optimal locations and supervision modes for all segments, according to the given criteria. Each resulting segment ill be assigned e actl one of these supervision modes. e denote a segmentation of the N to ens of corpus w_1^N into $M \le N$ segments b specif ing segment boundar mar ers $s_1^{M+1}=(s_1=1, s_2, \ldots, s_{M+1}=N+1)$. Setting a boundar mar er $s_i=a$ means that e put a segment boundar before the a-th ord to en (or the end-of-corpus mar er for $a=N+1$). Thus our corpus is segmented into to en se uences $[(w_{s_j}, \ldots, w_{s_{j+1}} \;\; 1)]_{j=1}^{M}$. The supervision modes assigned to each segment are denoted b m_j. e favor those segmentations that minimi e the cumulative value $\sum_{j=1}^{M}[u_{l,m_j}(w_{s_j}^{s_{j+1}})]$ for each criterion l. For an criterion here larger values are intuitivel better, e ip the sign before de ning $u_{l,m_j}(w_{s_j}^{s_{j+1}})$ to maintain consistenc (e.g. negative number of errors removed).

3.1 Multiple Criteria Optimization

n the case of a single criterion ($|L|=1$), e obtain a simple, single-ob ective unconstrained linear optimi ation problem, ef cientl solvable via d namic programming (Ter i and Tsaparas, 2006). o ever, in practice one usuall encounters several competing criteria, such as cost and utilit , and here e ill focus on this more realistic setting. e balance competing criteria b using one as an optimi ation ob ective, and the others as constraints.[1] Let crite-

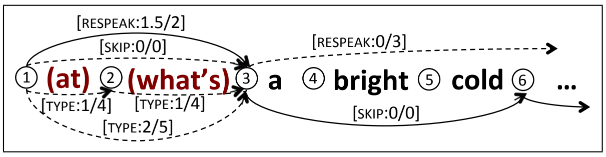

Figure 3: E cerpt of a segmentation graph for an e ample transcription tas similar to Figure 1 (some edges are omitted for readabilit). Edges are labeled ith their mode, predicted number of errors that can be removed, and necessar supervision time. A segmentation scheme might prefer solid edges over dashed ones in this e ample.

rion l be the optimi ation ob ective criterion, and let C_l denote the constraining constants for the criteria $l \in L \;\; l' = L \setminus \{l\}$. e state the optimi ation problem:

$$\min_{M; s_1^{M+1}; m_1^M} \sum_{j=1}^{M} \left[u_{l',m_j}\left(w_{s_j}^{s_{j+1}}\right) \right]$$

$$\text{s.t.} \sum_{j=1}^{M} \left[u_{l,m_j}\left(w_{s_j}^{s_{j+1}}\right) \right] \le C_l \quad (\forall l \in L \;\; l')$$

This constrained optimi ation problem is dif cult to solve. n fact, the P-hard multiple-choice napsac problem (Pisinger, 1994) corresponds to a special case of our problem in hich the number of segments is e ual to the number of to ens, impl ing that our more general problem is P-hard as ell.

n order to overcome this problem, e reformulate search for the optimal segmentation as a resource-constrained shortest path problem in a directed, ac clic multigraph. hile still not ef cientl solvable in theor , this problem is ell studied in domains such as vehicle routing and cre scheduling (rnich and esaulniers, 2005), and it is no n that in man practical situations the problem can be solved reasonabl ef cientl using integer linear programming rela ations (Toth and igo, 2001).

n our formalism, the set of nodes V represents the spaces bet een neighboring to ens, at hich the algorithm ma insert segment boundaries. A node ith inde i represents a segment brea before the i-th to en, and thus the se uence of the indices in a path directl corresponds to s_1^{M+1}. Edges E denote the grouping of to ens bet een the respective

[1] This approach is no n as the *bounded objective function method* in multi-ob ective optimi ation literature (arler and Arora, 2004). The ver popular *weighted sum method* merges criteria into a single ef cienc measure, but is problematic in our case because the number of supervised to ens is unspeci ed. nless the eights are carefull chosen, the algorithm might nd, e.g., the completel unsupervised or completel supervised segmentation to be most ef cient.

nodes into one segment. Edges are al a s directed from left to right, and labeled ith a supervision mode. n addition, each edge bet een nodes i and j is assigned $u_{l,k}(w_i^{j\ \ 1})$, the corresponding predicted value for each criterion $l \in L$ and supervision mode $k \in K$, indicating that the supervision mode of the j-th segment in a path directl corresponds to m_j.

Figure 3 sho s an e ample of hat the resulting graph ma loo li e. ur original optimi ation problem is no e uivalent to nding the shortest path bet een the rst and last nodes according to criterion l , hile obe ing the given resource constraints. According to a idel used formulation for the resource constrained shortest path problem, e can de ne E_{ij} as the set of competing edges bet een i and j, and e press this optimi ation problem ith the follo ing integer linear program (LP):

$$\min_{\mathbf{x}} \sum_{i,j\ \ V} \sum_{k\ \ E_{ij}} x_{ijk} u_{l',k}(s_i^{j\ \ 1}) \tag{1}$$

$$\text{s.t.} \sum_{i,j\ \ V} \sum_{k\ \ E_{ij}} x_{ijk} u_{l,k}(s_i^{j\ \ 1}) \leq C_l \qquad (\forall l \in L\ \ l') \tag{2}$$

$$\sum_{\substack{i\ \ V \\ k\ \ E_{ij}}} x_{ijk} = \sum_{\substack{i\ \ V \\ k\ \ E_{ij}}} x_{jik} \qquad (\forall j \in V \backslash \{1, n\}) \tag{3}$$

$$\sum_{\substack{j\ \ V \\ k\ \ E_{1j}}} x_{1jk} = 1 \tag{4}$$

$$\sum_{\substack{i\ \ V \\ k\ \ E_{in}}} x_{ink} = 1 \tag{5}$$

$$x_{ijk} \in \{0, 1\} \qquad (\forall x_{ijk} \in \mathbf{x}) \tag{6}$$

The variables $\mathbf{x}=\{x_{ijk} | i, j \in V, k \in E_{ij}\}$ denote the activation of the k th edge bet een nodes i and j. The shortest path according to the minimi ation ob ective (1), that still meets the resource constraints for the speci ed criteria (2), is to be computed. The degree constraints (3,4,5) specif that all but the rst and last nodes must have as man incoming as outgoing edges, hile the rst node must have e actl one outgoing, and the last node e actl one incoming edge. Finall , the integralit condition (6) forces all edges to be either full activated or full deactivated. The outlined problem formulation can solved directl b using off-the-shelf LP solvers, here e emplo R (urobi ptimi ation, 2012).

3.2 Heuristics for Approximation

n general, edges are inserted for ever supervision mode bet een ever combination of t o nodes. The search space can be constrained b removing some of these edges to increase ef cienc . n this stud , e onl consider edges spanning at most 20 to ens.

For cases in hich larger corpora are to be annotated, or hen the acceptable dela for delivering results is small, a suitable segmentation can be found appro imatel . The easiest a ould be to partition the corpus, e.g. according to its individual documents, divide the budget constraints evenl across all partitions, and then segment each partition independentl . ore sophisticated methods might appro imate the Pareto front for each partition, and distribute the budgets in an intelligent a .

4 User Modeling

hile the proposed frame or is able to optimi e the segmentation ith respect to each criterion, it also rests upon the assumption that e can provide user models $u_{l,k}(w_i^{j\ \ 1})$ that accuratel evaluate ever segment according to the speci ed criteria and supervision modes. n this section, e discuss our strategies for estimating three conceivable criteria: annotation cost, correction of errors, and improvement of a classi er.

4.1 Annotation Cost Modeling

odeling cost re uires solving a regression problem from features of a candidate segment to annotation cost, for e ample in terms of supervision time. Appropriate input features depend on the tas , but should include notions of comple it (e.g. a con dence measure) and length of the segment, as both are e pected to strongl in uence supervision time.

e propose using aussian process (P) regression for cost prediction, a start-of-the-art nonparametric a esian regression techni ue (Rasmussen and illiams, 2006)[2]. As reported on a similar tas b Cohn and Specia (2013), and con rmed b our preliminar e periments, P regression signi cantl outperforms popular techni ues such as sup-

[2] Code available at http:// .gaussianprocess.org/gpml/

port vector regression and least-s uares linear regression. e also follo their settings for P, emplo ing P regression ith a s uared e ponential ernel ith automatic relevance determination. epending on the number of users and amount of training data available for each user, models ma be trained separatel for each user (as e do here), or in a combined fashion via multi-tas learning as proposed b Cohn and Specia (2013).

t is also crucial for the predictions to be reliable throughout the hole relevant space of segments. f the cost of certain t pes of segments is s stematicall underpredicted, the segmentation algorithm might be misled to prefer these, possibl a large number of times.[3] An effective tric to prevent such underpredictions is to predict the log time instead of the actual time. n this a , errors in the critical lo end are penali ed more strongl , and the time can never become negative.

4.2 Error Correction Modeling

As one utilit measure, e can use the number of errors corrected, a useful measure for post editing tas s over automaticall produced annotations. n order to measure ho man errors can be removed b supervising a particular segment, e must estimate both ho man errors are in the automatic annotation, and ho reliabl a human can remove these for a given supervision mode.

ost machine learning techni ues can estimate con dence scores in the form of posterior probabilities. To estimate the number of errors, e can sum over one minus the posterior for all to ens, hich estimates the amming distance from the reference annotation. This measure is appropriate for tas s in hich the number of to ens is ed in advance (e.g. a part-of-speech estimation tas), and a reasonable appro imation for tas s in hich the number of to ens is not no n in advance (e.g. speech transcription, cf. Section 5.1.1).

Predicting the particular to ens at hich a human ill ma e a mista e is no n to be a dif cult tas (lson and lson, 1990), but a simplif ing constant human error rate can still be useful. For e ample, in the tas from Section 2, e ma suspect a certain number of errors in a transcript segment, and predict, sa , 95 of those errors to be removed via t ping, but onl 85 via respea ing.

4.3 Classifier Improvement Modeling

Another reasonable utilit measure is accurac of a classi er trained on the data e choose to annotate in an active learning frame or . Con dence scores have been found useful for ran ing particular to ens ith regards to ho much the ill improve a classi er (Settles, 2008). ere, e ma similarl score segment utilit as the sum of its to en con dences, although care must be ta en to normali e and calibrate the to en con dences to be linearl comparable before doing so. hile the resulting utilit score has no interpretation in absolute terms, it can still be used as an optimi ation ob ective (cf. Section 5.2.1).

5 Experiments

n this section, e present e perimental results e amining the effectiveness of the proposed method over t o tas s: speech transcription and apanese ord segmentation.[4]

5.1 Speech Transcription Experiments

Accurate speech transcripts are a much-demanded LP product, useful b themselves, as training material for ASR, or as input for follo -up tas s li e speech translation. ith recognition accuracies plateauing, manuall correcting (post editing) automatic speech transcripts has become popular. Common approaches are to identif ords (Sanche - Cortina et al., 2012) or (sub-)sentences (Sperber et al., 2013) of lo con dence, and have a human editor correct these.

5.1.1 Experimental Setup

e conduct a user stud in hich participants post-edited speech transcripts, given a ed goal ord error rate. The transcription setup as such that the transcriber could see the ASR transcript of parts before and after the segment that he as editing, providing conte t if needed. hen imprecise time alignment resulted in segment brea s that ere

[3]For instance, consider a model that predicts ell for segments of medium si e or longer, but underpredicts the supervision time of single-to en segments. This ma lead the segmentation algorithm to put ever to en into its o n segment, hich is clearl undesirable.

[4]Soft are and e perimental data can be do nloaded from http:// .msperber.com/research/tacl-segmentation/

slightl off, as happened occasionall , that conte t helped guess hat as said. The segment itself as transcribed from scratch, as opposed to editing the ASR transcript; besides being arguabl more ef - cient hen the ASR transcript contains man mis- ta es (an o et al., 2006; A ita et al., 2009), prelim- inar e periments also sho ed that supervision time is far easier to predict this a . Figure 4 illustrates hat the setup loo ed li e.

e used a self-developed transcription tool to conduct e periments. t presents our computed seg- ments one b one, allo s convenient input and pla - bac via e board shortcuts, and logs user interac- tions ith their time stamps. A selection of TE tal s[5] (English tal s on technolog , entertainment, and design) served as e perimental data. hile some of these tal s contain argon such as medi- cal terms, the are presented b s illed spea ers, ma ing them comparabl eas to understand. nitial transcripts ere created using the anus recognition tool it (Soltau et al., 2001) ith a standard, TE - optimi ed setup. e used confusion net or s for decoding and obtaining con dence scores.

For reasons of simplicit , and better compara- bilit to our baseline, e restricted our e periment to t o supervision modes: T PE and S P. e conducted e periments ith 3 participants, 1 ith several ears of e perience in transcription, 2 ith none. Each participant received an e planation on the transcription guidelines, and a short hands-on training to learn to use our tool. e t, the tran- scribed a balanced selection of 200 segments of var ing length and ualit in random order. This data as used to train the user models.

Finall , each participant transcribed another 2 TE tal s, ith ord error rate (ER) 19.96 (predicted: 22.33). e set a target (predicted) ER of 15 as our optimi ation constraint,[6] and minimi e the predicted supervision time as our ob- ective function. oth TE tal s ere transcribed once using the baseline strateg , and once using the proposed strateg . The order of both strategies as reversed bet een tal s, to minimi e learning bias due to transcribing each tal t ice.

The baseline strateg as adopted according to Sperber et al. (2013): e segmented the tal into natural, subsentential units, using atusov et al. (2006) s segmenter, hich e tuned to reproduce the TE subtitle segmentation, producing a mean segment length of 8.6 ords. Segments ere added in order of increasing average ord con dence, until the user model predicted a ER$<15\%$. The second segmentation strateg as the proposed method, similarl ith a resource constraint of ER$<15\%$.

Supervision time as predicted via P regres- sion (cf. Section 4.1), using segment length, au- dio duration, and mean con dence as input features. The output variable as assumed sub ect to addi- tive aussian noise ith ero mean, a variance of 5 seconds as chosen empiricall to minimi e the mean s uared error. tilit prediction (cf. Section 4.2) as based on posterior scores obtained from the confusion net or s. e found it important to calibrate them, as the posteriors ere overcon dent especiall in the upper range. To do so, e automat- icall transcribed a development set of TE data, grouped the recogni ed ords into buc ets accord- ing to their posteriors, and determined the average number of errors per ord in each buc et from an alignment ith the reference transcript. The map- ping from average posterior to average number of errors as estimated via P regression. The result as summed over all to ens, and multiplied b a constant human con dence, separatel determined for each participant.[7]

5.1.2 Simulation Results

To conve a better understanding of the poten- tial gains afforded b our method, e rst present a simulated e periment. e assume a transcriber ho ma es no mista es, and needs e actl the amount of time predicted b a user model trained on the data of a randoml selected participant. e compare three scenarios: A baseline simulation, in hich the base- line segments are transcribed in ascending order of con dence; a simulation using the proposed method, in hich e change the ER constraint in small in- crements; nall , an oracle simulation, hich uses

[5] .ted.com

[6] epending on the level of accurac re uired b our nal application, this target ma be set lo er or higher.

[7] ore elaborate methods for ER estimation e ist, such as b ga a et al. (2013), but if our method achieves improve- ments using simple amming distance, incorporating more so- phisticated measures ill li el achieve similar, or even better accurac .

(3) S P: nineteen fort si until toda ou see the green
(4) T PE: <annotator t pes: is the traditional >
(5) S P: nterstate con ict
(6) T PE: <annotator t pes: the ones e used to >
(7) S P: . . .

Figure 4: Result of our segmentation method (e cerpt). T PE segments are displa ed empt and should be transcribed from scratch. For S P segments, the ASR transcript is displa ed to provide conte t. hen annotating a segment, the corresponding audio is pla ed bac .

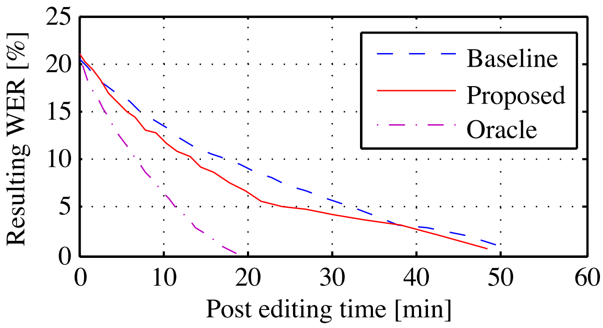

Figure 5: Simulation of post editing on e ample TE tal . The proposed method reduces the ER considerabl faster than the baseline at rst, later both converge. The much superior oracle simulation indicates room for further improvement.

the proposed method, but uses a utilit model that no s the actual number of errors in each segment. For each supervised segment, e simpl replace the ASR output ith the reference, and measure the resulting ER.

Figure 5 sho s the simulation on an e ample TE tal , based on an initial transcript ith 21.9 ER. The proposed method is able to reduce the ER faster than the baseline, up to a certain point here the converge. The oracle simulation is even faster, indicating room for improvement through better con dence scores.

5.1.3 User Study Results

Table 1 sho s the results of the user stud . First, e note that the ER estimation b our utilit model as off b about 2.5 : hile the predicted improvement in ER as from 22.33 to 15.0 , the actual improvement as from 19.96 to about 12.5 . The actual resulting ER as consistent

Participant	Baseline		Proposed	
	ER	Time	ER	Time
P_1	12.26	44:05	12.18	33:01
P_2	12.75	36:19	12.77	29:54
P_3	12.70	52:42	12.50	37:57
A	12.57	44:22	12.48	33:37

Table 1: Transcription tas results. For each user, the resulting ER after supervision is sho n, along ith the time min the needed. The unsupervised ER as 19.96 .

across all users, and e observe strong, consistent reductions in supervision time for all participants. Prediction of the necessar supervision time as accurate: Averaged over participants, 45:41 minutes ere predicted for the baseline, 44:22 minutes measured. For the proposed method, 32:11 minutes ere predicted, 33:37 minutes measured. n average, participants removed 6.68 errors per minute using the baseline, and 8.93 errors per minute using the proposed method, a speed-up of 25.2 .

ote that predicted and measured values are not strictl comparable: n the e periments, to provide a fair comparison participants transcribed the same tal s t ice (once using baseline, once the proposed method, in alternating order), resulting in a noticeable learning effect. The user model, on the other hand, is trained to predict the case in hich a transcriber conducts onl one transcription pass.

As an interesting nding, ithout being informed about the order of baseline and proposed method, participants reported that transcribing according to the proposed segmentation seemed harder, as the found the baseline segmentation more linguisticall reasonable. o ever, this perceived increase in dif cult did not sho in ef cienc numbers.

5.2 Japanese Word Segmentation Experiments

ord segmentation is the rst step in LP for languages that are commonl ritten ithout ord boundaries, such as apanese and Chinese. e appl our method to a tas in hich e domain-adapt a ord segmentation classi er via active learning. n this e periment, participants annotated hether or not a ord boundar occurred at certain positions in a apanese sentence. The to ens to be grouped into segments are positions bet een ad acent characters.

5.2.1 Experimental Setup

eubig et al. (2011) have proposed a point ise method for apanese ord segmentation that can be trained using partiall annotated sentences, hich ma es it attractive in combination ith active learning, as ell as our segmentation method. The authors released their method as a soft are pac age Tea that e emplo ed in this user stud . e used Tea s active learning domain adaptation tool it[8] as a baseline.

For data, e used the alanced Corpus of Contemporar ritten apanese (CC), created b ae a a (2008), ith the *internet Q&A subcorpus* as in-domain data, and the *whitepaper* subcorpus as bac ground data, a domain adaptation scenario. Sentences ere dra n from the in-domain corpus, and the manuall annotated data as then used to train Tea, along ith the pre-annotated bac ground data. The goal (ob ective function) as to improve Tea s classi cation accurac on an in-domain test set, given a constrained time budget of 30 minutes. There ere again 2 supervision modes: A TATE and S P. ote that this is essentiall a batch active learning setup ith onl one iteration.

e conducted e periments ith one e pert ith several ears of e perience ith apanese ord segmentation annotation, and three non-e pert native spea ers ith no prior e perience. apanese ord segmentation is not a trivial tas , so e provided non-e perts ith training, including e planation of the segmentation standard, a supervised test ith immediate feedbac and e planations, and hands-on training to get used to the annotation soft are.

Supervision time as predicted via P regression (cf. Section 4.1), using the segment length and mean con dence as input features. As before, the output variable as assumed sub ect to additive aussian noise ith ero mean and 5 seconds variance. To obtain training data for these models, each participant annotated about 500 e ample instances, dra n from the adaptation corpus, grouped into segments and balanced regarding segment length and dif cult .

For utilit modeling (cf. Section 4.3), e rst normali ed Tea s con dence scores, hich are given in terms of S margin, using a sigmoid function (Platt, 1999). The normali ation parameter as selected so that the mean con dence on a development set corresponded to the actual classi er accurac .

e derive our measure of classi er improvement for correcting a segment b summing over one minus the calibrated con dence for each of its to ens. To anal e ho ell this measure describes the actual training utilit , e trained Tea using the bac ground data plus dis oint groups of 100 in-domain instances ith similar probabilities and measured the achieved reduction of prediction errors. The correlation bet een each group s mean utilit and the achieved error reduction as 0.87. ote that e ignore the deca ing returns usuall observed as more data is added to the training set. Also, e did not attempt to model user errors. Emplo ing a constant base error rate, as in the transcription scenario, ould change segment utilities onl b a constant factor, ithout changing the resulting segmentation.

After creating the user models, e conducted the main e periment, in hich each participant annotated data that as selected from a pool of 1000 in-domain sentences using t o strategies. The rst, baseline strateg as as proposed b eubig et al. (2011). ueries are those instances ith the lo est con dence scores. Each uer is then e tended to the left and right, until a ord boundar is predicted. This strateg follo s similar reasoning as as the premise to this paper: To decide hether or not a position in a te t corresponds to a ord boundar , the annotator has to ac uire surrounding conte t information. This conte t ac uisition is relativel time consuming, so he might as ell label the surrounding instances ith little additional effort. The second strateg as our proposed, more principled approach. ueries of both methods ere shuf ed to minimi e bias due to learning effects. Finall , e trained Tea using the results of both methods, and compared the achieved classi er improvement and supervision times.

5.2.2 User Study Results

Table 2 summari es the results of our e periment. t sho s that the annotations b each participant resulted in a better classi er for the proposed method than the baseline, but also too up considerabl more time, a less clear improvement than for the transcription tas . n fact, the total error for time predictions as as high as 12.5 on average,

[8] http:// .phontron.com/ tea/active.html

Participant	Baseline		Proposed	
	Time	Acc.	Time	Acc.
E pert	25:50	96.17	32:45	96.55
onE p$_1$	22:05	95.79	26:44	95.98
onE p$_2$	23:37	96.15	31:28	96.21
onE p$_3$	25:23	96.38	33:36	96.45

Table 2: ord segmentation tas results, for our e - pert and 3 non-e pert participants. For each participant, the resulting classi er accurac after supervision is sho n, along ith the time min the needed. The unsupervised accurac as 95.14 .

here the baseline method tended ta e less time than predicted, the proposed method more time. This is in contrast to a much lo er total error (ithin 1) hen cross-validating our user model training data. This is li el due to the fact that the data for training the user model as selected in a balanced manner, as opposed to selecting dif cult e amples, as our method is prone to do. Thus, e ma e pect much better predictions hen selecting user model training data that is more similar to the test case.

Plotting classi er accurac over annotation time dra s a clearer picture. Let us rst anal e the results for the e pert annotator. Figure 6 (E.1) sho s that the proposed method resulted in consistentl better results, indicating that time predictions ere still effective. ote that this comparison ma put the proposed method at a slight disadvantage b comparing intermediate results despite optimi ing globall .

For the non-e perts, the improvement over the baseline is less consistent, as can be seen in Figure 6 (.1) for one representative. According to our anal sis, this can be e plained b t o factors: (1) The non-e perts annotation error (6.5 on average) as much higher than the e pert s (2.7), resulting in a some hat irregular classi er learning curve. (2) The variance in annotation time per segment as consistentl higher for the non-e perts than the e pert, indicated b an average per-segment prediction error of 71 vs. 58 relative to the mean actual value, respectivel . nformall spea ing, non-e perts made more mista es, and ere more strongl in uenced b the dif cult of a particular segment (hich as higher on average ith the proposed method, as indicated b a

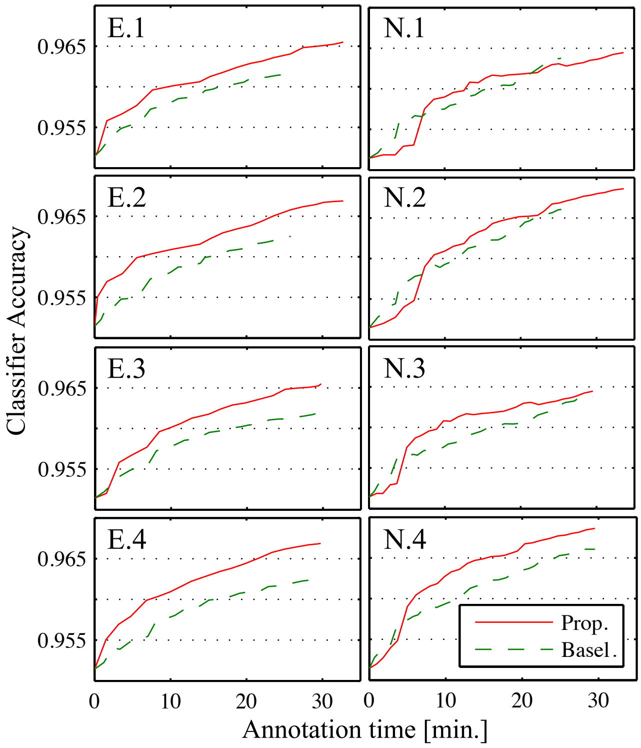

Figure 6: Classi er improvement over time, depicted for the e pert (E) and a non-e pert ( ). The graphs sho numbers based on (1) actual annotations and user models as in Sections 4.1 and 4.3, (2) error-free annotations, (3) measured times replaced b predicted times, and (4) both reference annotations and replaced time predictions.

lo er average con dence).[9]

n Figures 6 (2-4) e present a simulation e periment in hich e rst pretend as if annotators made no mista es, then as if the needed e actl as much time as predicted for each segment, and then both. This cheating e periment or s in favor of the proposed method, especiall for the non-e pert. e ma conclude that our segmentation approach is effective for the ord segmentation tas , but re uires more accurate time predictions. etter user models ill certainl help, although for the presented scenario our method ma be most useful for an e pert annotator.

[9] ote that the non-e pert in the gure annotated much faster than the e pert, hich e plains the comparable classi cation result despite ma ing more annotation errors. This is in contrast to the other non-e perts, ho ere slo er.

5.3 Computational Efficiency

Since our segmentation algorithm does not guarantee pol nomial runtime, computational ef cienc as a concern, but did not turn out problematic. n a consumer laptop, the solver produced segmentations ithin a fe seconds for a single document containing several thousand to ens, and ithin hours for corpora consisting of several do en documents. Runtime increased roughl uadraticall ith respect to the number of segmented to ens. e feel that this is acceptable, considering that the time needed for human supervision ill li el dominate the computation time, and reasonable appro imations can be made as noted in Section 3.2.

6 Relation to Prior Work

Ef cient supervision strategies have been studied across a variet of LP-related research areas, and received increasing attention in recent ears. E amples include post editing for speech recognition (Sanche -Cortina et al., 2012), interactive machine translation (on ale -Rubio et al., 2010), active learning for machine translation (affari et al., 2009; on ale -Rubio et al., 2011) and man other LP tas s (lsson, 2009), to name but a fe studies.

t has also been recogni ed b the active learning communit that correcting the most useful parts rst is often not optimal in terms of ef cienc , since these parts tend to be the most dif cult to manuall annotate (Settles et al., 2008). The authors advocate the use of a user model to predict the supervision effort, and select the instances ith best bang-for-the-buc . This prediction of supervision effort as successful, and as further re ned in other LP-related studies (Tomane et al., 2010; Specia, 2011; Cohn and Specia, 2013). ur approach to user modeling using P regression is inspired b the latter.

ost studies on user models consider onl supervision effort, hile neglecting the accurac of human annotations. The vie on humans as a perfect oracle has been critici ed (onme and Carbonell, 2008), since human errors are common and can negativel affect supervision utilit . Research on human-computer-interaction has identi ed the modeling of human errors as ver dif cult (lson and lson, 1990), depending on factors such as user e perience, cognitive load, user interface design, and fatigue. evertheless, even the simple error model used in our post editing tas as effective.

The active learning communit has addressed the problem of balancing utilit and cost in some more detail. The previousl reported bang-for-the-buc approach is a ver simple, greed approach to combine both into one measure. A more theoreticall founded scalar optimi ation ob ective is the net bene t (utilit minus costs) as proposed b i a anarasimhan and rauman (2009), but unfortunatel is restricted to applications here both can be e pressed in terms of the same monetar unit. i a anarasimhan et al. (2010) and onme and Carbonell (2008) use a more practical approach that speci es a constrained optimi ation problem b allo ing onl a limited time budget for supervision. ur approach is a generali ation thereof and allo s either specif ing an upper bound on the predicted cost, or a lo er bound on the predicted utilit .

The main novelt of our presented approach is the e plicit modeling and selection of segments of various si es, such that annotation ef cienc is optimi ed according to the speci ed constraints. hile some or s (Sassano and urohashi, 2010; eubig et al., 2011) have proposed using subsentential segments, e are not a are of an previous or that e plicitl optimi es that segmentation.

7 Conclusion

e presented a method that can effectivel choose a segmentation of a language corpus that optimi es supervision ef cienc , considering not onl the actual usefulness of each segment, but also the annotation cost. e reported noticeable improvements over strong baselines in t o user studies. Future user e periments ith more participants ould be desirable to verif our observations, and allo further anal sis of different factors such as annotator e pertise. Also, future research ma improve the user modeling, hich ill be bene cial for our method.

Acknowledgments

The research leading to these results has received funding from the European nion Seventh Frame or Programme (FP7/2007-2013) under grant agreement n 287658 ridges Across the Language ivide (E - R E).

References

u a A ita, asato imura, and Tatsu a a ahara. 2009. Automatic Transcription S stem for eetings of the apanese ational Congress. n *Interspeech*, pages 84–87, righton, .

Trevor Cohn and Lucia Specia. 2013. odelling Annotator ias ith ulti-tas aussian Processes: An Application to achine Translation ualit Estimation. n *Association for Computational Linguistics Conference (ACL)*, So a, ulgaria.

Pinar onme and aime Carbonell. 2008. Proactive Learning : Cost-Sensitive Active Learning ith ultiple mperfect racles. n *Conference on Information and Knowledge Management (CIKM)*, pages 619–628, apa alle , CA, SA.

esus on ale -Rubio, aniel rti - art ne , and Francisco Casacuberta. 2010. alancing ser Effort and Translation Error in nteractive achine Translation ia Con dence easures. n *Association for Computational Linguistics Conference (ACL), Short Papers Track*, pages 173–177, ppsala, S eden.

esus on ale -Rubio, aniel rti - art ne , and Francisco Casacuberta. 2011. An active learning scenario for interactive machine translation. n *International Conference on Multimodal Interfaces (ICMI)*, pages 197–200, Alicante, Spain.

urobi ptimi ation. 2012. urobi ptimi er Reference anual.

holamre a affari, a im Ro , and Anoop Sar ar. 2009. Active Learning for Statistical Phrase-based achine Translation. n *North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference (NAACL-HLT)*, pages 415–423, oulder, C , SA.

Stefan rnich and u esaulniers. 2005. Shortest Path Problems ith Resource Constraints. n *Column Generation*, pages 33–65. Springer S.

i uo ae a a. 2008. alanced Corpus of Contemporar ritten apanese. n *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 101–102, derabad, ndia.

R. Timoth arler and asbir S. Arora. 2004. Surve of multi-ob ective optimi ation methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, April.

Evgen atusov, Arne auser, and ermann e . 2006. Automatic Sentence Segmentation and Punctuation Prediction for Spo en Language Translation. n *International Workshop on Spoken Language Translation (IWSLT)*, pages 158–165, oto, apan.

iroa i an o, u a A ita, and Tatsu a a ahara. 2006. Computer Assisted Speech Transcription S stem for Ef cient Speech Archive. n *Western Pacific Acoustics Conference (WESPAC)*, Seoul, orea.

raham eubig, osu e a ata, and Shinsu e ori. 2011. Point ise Prediction for Robust , Adaptable apanese orphological Anal sis. n *Association for Computational Linguistics: Human Language Technologies Conference (ACL-HLT)*, pages 529–533, Portland, R, SA.

Atsunori ga a, Ta aa i ori, and Atsushi a amura. 2013. iscriminative Recognition Rate Estimation For - est List and ts Application To - est Rescoring. n *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6832–6836, ancouver, Canada.

udith Reitman lson and ar lson. 1990. The ro th of Cognitive odeling in uman-Computer nteraction Since S. *Human-Computer Interaction*, 5(2):221–265, une.

Fredri lsson. 2009. A literature surve of active machine learning in the conte t of natural language processing. Technical report, S CS S eden.

avid Pisinger. 1994. A inimal Algorithm for the ultiple-Choice napsac Problem. *European Journal of Operational Research*, 83(2):394–410.

ohn C. Platt. 1999. Probabilistic utputs for Support ector achines and Comparisons to Regulari ed Li elihood ethods. n *Advances in Large Margin Classifiers*, pages 61–74. T Press.

Carl E. Rasmussen and Christopher . . illiams. 2006. *Gaussian Processes for Machine Learning*. T Press, Cambridge, A, SA.

saias Sanche -Cortina, icolas Serrano, Alberto Sanchis, and Alfons uan. 2012. A protot pe for nteractive Speech Transcription alancing Error and Supervision Effort. n *International Conference on Intelligent User Interfaces (IUI)*, pages 325–326, Lisbon, Portugal.

anabu Sassano and Sadao urohashi. 2010. sing Smaller Constituents Rather Than Sentences in Active Learning for apanese ependenc Parsing. n *Association for Computational Linguistics Conference (ACL)*, pages 356–365, ppsala, S eden.

urr Settles, ar Craven, and Le is Friedland. 2008. Active Learning ith Real Annotation Costs. n *Neural Information Processing Systems Conference (NIPS) - Workshop on Cost-Sensitive Learning*, La e Tahoe, , nited States.

urr Settles. 2008. An Anal sis of Active Learning Strategies for Se uence Labeling Tas s. n *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1070–1079, onolulu, SA.

agen Soltau, Florian et e, Christian Fugen, and Ale aibel. 2001. A ne-Pass ecoder ased on Pol morphic Linguistic Conte t Assignment. n *Automatic Speech Recognition and Understanding Work-*

shop (ASRU), pages 214–217, adonna di Campiglio, tal .

Lucia Specia. 2011. E ploiting b ective Annotations for easuring Translation Post-editing Effort. n *Conference of the European Association for Machine Translation (EAMT)*, pages 73–80, ice, France.

atthias Sperber, raham eubig, Christian Fugen, Satoshi a amura, and Ale aibel. 2013. Ef cient Speech Transcription Through Respea ing. n *Interspeech*, pages 1087–1091, L on, France.

ernhard Suhm, rad ers, and Ale aibel. 2001. ultimodal error correction for speech user interfaces. *Transactions on Computer-Human Interaction*, 8(1):60–98.

Evimaria Ter i and Pana iotis Tsaparas. 2006. Ef cient algorithms for se uence segmentation. n *SIAM Conference on Data Mining (SDM)*, ethesda, , SA.

atrin Tomane and do ahn. 2009. Semi-Supervised Active Learning for Se uence Labeling. n *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1039–1047, Singapore.

atrin Tomane , do ahn, and Steffen Lohmann. 2010. A Cognitive Cost odel of Annotations ased on E e-Trac ing ata. n *Association for Computational Linguistics Conference (ACL)*, pages 1158–1167, ppsala, S eden.

Paolo Toth and aniele igo. 2001. *The Vehicle Routing Problem*. Societ for ndustrial Applied athematics (S A), Philadelphia.

Sudheendra i a anarasimhan and risten rauman. 2009. hats t oing to Cost ou : Predicting Effort vs. nformativeness for ulti-Label mage Annotations. n *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2262–2269, iami each, FL, SA.

Sudheendra i a anarasimhan, Pratee ain, and risten rauman. 2010. Far-sighted active learning on a budget for image and video recognition. n *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3035–3042, San Francisco, CA, SA, une.