WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Languages

Nathan Green[1,2,*] **Paul Breimyer**[1,2,*] **Vinay Kumar**[1] **Nagiza F. Samatova**[1,2,∓]

[1]**North Carolina State University**
890 Oval Drive
Raleigh, NC 27695

[2]**Oak Ridge National Laboratory**
1 Bethel Valley Rd
Oak Ridge, TN 37831

Abstract

Annotated corpora are sets of structured text used to enable Natural Language Processing (NLP) tasks. Annotations may include tagged parts-of-speech, semantic concepts assigned to phrases, or semantic relationships between these concepts in text. Building annotated corpora is labor-intensive and presents a major obstacle to advancing machine translators, named entity recognizers (NER), part-of-speech taggers, etc. Annotated corpora are specialized for a particular language or NLP task. Hence, a majority of the world's 6000+ languages lack NLP resources, and therefore remain *minority*, or *under-resourced*, languages in modern language technologies.

In this paper we present WebBANC, a framework for Building Annotated NLP Corpora from user annotations on the Web. With WebBANC, a casual user can annotate parts of HTML or PDF text on any website and associate the text with semantic concepts specific to an NLP task. User annotations are combined by WebBANC to produce annotated corpora potentially comparable in diversity to corpora in English, minority languages, and human generated categories, such as those on Yahoo.com, with an average precision and recall of 0.80, which is comparable to automated NER tools on the CoNLL benchmark.

* Both authors contributed equally
∓ Corresponding author: samatovan@ornl.gov

1 Introduction

The Web is the holy grail of linguistic data (Rayson et al., 2006). It has recently gained popularity as a resource for *minority* (Ghani and Mladenic, 2001), or *under-resourced*, languages that lack automatic Natural Language Processing (NLP) resources, even from the Basic Language Resource Kit (BLARK) (Krauwer, 2003). "Web as Corpus" has been especially valuable for constructing text corpora from the Web for these languages (Scannell, 2007; Baroni and Bernardini, 2004). Language specific corpora are useful for many language technology applications, including named entity recognition, machine translation, spelling correction, and machine-readable dictionaries. The An Crúbadán Project, for example, has succeeded in creating corpora for more than 400 of the world's 6000+ languages by web crawling. With a few exceptions, most of the 400+ corpora, however, lack any linguistic annotations due to the limitations of the annotation tools (Rayson et al., 2006).

In spite of the many documented advantages of linguistically annotated data over raw data (Mair, 2005), annotated corpora are quite sparse. The majority of previous work on corpus annotation has utilized manual coding by linguistic experts, automated software tagging systems, and semi-automatic combinations of the two approaches. Uren et al. provide a comprehensive survey of existing semantic annotation tools, including some community-driven projects (2006). While yielding high quality and enormous value, manual corpus annotation is both tedious and time-consuming. For example, the GENIA corpus contains 9,372 sentences, curated by five part-time annotators, one senior coordinator, and one junior coordinator over 1.5 years (Kim et al., 2008). In contrast, software tagging systems, such as those for annotating web corpora are automatic and fast,

but primarily exist for *majority* languages.

For *minority* languages, however, few automated corpora annotation systems exist and different approaches are needed. In this paper, we hypothesize that the Web, coupled with web user community efforts, represent a paradigm shift in annotated corpora construction. We extend the concept of community-based web content creation, such as Wikipedia (Zesch et al., 2007), by assuming that websites, especially frequently visited ones, present an ideal platform for large-scale community-level annotations for NLP tasks. We also argue that if given an opportunity to link annotations with semantic concepts, such as those represented in the form of ontologies, the web community can potentially create semantically-rich annotated corpora at an unprecedented scale.

The actual impact of web user annotated corpora creation remains to be seen, but the potential benefits of such a framework are manifold. It may reduce the time required to create annotated corpora for NLP tasks potentially from months to days. For NER tasks, for example, commercial applications currently support a handful of entities. For instance, NetOwl Extractor is a commercial application that supports seven entity types and seventy subtypes, including people, organizations, places, etc. The lack of entity breadth is explained by the intense human-labor required for entity type development.

A framework could potentially enable building semantically richer and larger corpora by supporting any ontology, which would allow researchers to introduce new levels of semantic richness into corpora. For example, the Gene Ontology (GO) (Ashburner et al., 2000) contains over 100,000 biological concepts that can enrich annotations and the correspondingly generated corpora.

A web user annotation framework may also enable automatic processing of minority languages by supporting minority corpora generation. The Open American National Corpus (OANC) (Ide and Macleod, 2001) is a major initiative meant to parallel the British National Corpus (Burnard, 1995), which contains over 100 million words. Minority languages do not enjoy the same support as American and British English, and it is unlikely that similar scale corpora will be generated for minority languages. The WebBANC framework can potentially enable annotated corpora generation of many less common domains, such as minority languages, by distributing the annotation effort over many users.

2 WebBANC Framework

We introduce a framework that leverages user annotations on the *Web* to *B*uild *A*nnotated *N*LP *C*orpora (WebBANC). We show that given such a framework, user annotations of commonly visited websites may contain enough linguistically diverse text to create sufficiently diverse corpora for various NLP tasks. To evaluate the results, we compare corpora created from the most visited websites to the human organized categories on Yahoo.com, and to commonly used corpora such as the OANC (Ide and Macleod, 2001), a freely available massive collection of American English texts with over fourteen million words.

We also compare the corpora against a minority language corpus generated from the Icelandic Frequency Dictionary (IFD) (Pind et al., 1991), a balanced corpus including Icelandic Fiction, Translated Fiction, and other categories compiled from text fragments written between 1980-1989 (Helgadttir, 2004). We show, through large-scale simulation, that aggregate user annotations covering approximately 50% of the words in the top 100 most visited websites can generate corpora that represent 35%-70% of the diversity of these corpora at 70%-90% precision. Small-scale user studies show that the average precision and recall for English named entity recognition (NER) tasks are comparable with those achieved by more than a dozen automatic NER tools when tested against the widely accepted CoNLL benchmark (Sang and Meulder, 2003).

2.1 Requirements

To be successful, a distributed free-text annotation framework must support annotations of most webpages that the *layman* user regularly encounters on the Web. For this reason, the framework should allow users to annotate both PDF and HTML documents, including pages built by underlying technologies that display HTML, such as PHP. Building corpora using distributed annotations should adhere to standards in the machine learning community, such as those proposed by the W3C, to enable standardized interfaces between clients and the framework. These standards may include the *Resource Description Framework* (RDF) (Klyne and Carroll, 2004) to communicate between the

web browser (the client) and the annotation manager (the server) and *XPointers* (DeRose et al., 1998) to locate text in HTML documents, allowing users to annotate any text on a webpage.

The framework should also provide easy-to-use annotation plug-ins for diverse web browsers with intuitive Graphical User Interfaces, potentially customized for individual NLP tasks. A simple drag-and-drop or right-mouse-click-and-select interface to choose a semantic concept, such as person or location for a highlighted word or phrase on the webpage, can serve as an example interface for NER tasks. Designing a simple and functional interface for different NLP tasks, such as entity relationships, may not be trivial.

A major issue for future minority language NLP developments is the need to generate and use *consistent* annotations (Leitner and Valencia, 2008). The framework should use standard semantic tags and allow user communities to supply their own standards; various scenarios are described below.

The framework should allow users to supply their own semantic tags for annotations. However, maintaining consistency may be quite difficult and may ultimately restrict the resulting annotated corpora uses for NLP tools.

The framework should permit users to choose semantic concepts and/or relationships from collections of controlled vocabularies, synonymous sets, and standard ontologies. Ontologies are formal representations of a set of domain concepts and the relationships between those concepts, and can provide a natural and standard hierarchy to tag a document. The W3C Web Ontology Language (OWL) (Bechhofer et al., 2004) is a standard for well-structured representations. Different domains have developed domain-specific ontologies, such as the Gene Ontology (GO) terms in Biology (Ashburner et al., 2000), but they may be too complex and require some adaptation to facilitate use by layman users, as well as domain experts. While the framework should allow users to select from a set of default ontologies, individual users and user communities should be free to create and integrate their own ontologies into the framework.

The framework should support semi-automated NLP tools or models to pre-annotate possibly relevant terms using existing NLP tools. The tools should use standard collections of semantic tags and offer the tagged annotations to users for validation via easy-to-use graphical interfaces. Semi-automated predictive models exist for some NLP tasks, such as part-of-speech and NER (Sang and Meulder, 2003). These models can be leveraged by the framework to validate manual annotations and may help identify poor annotations. Incorporating both ontologies and automated NLP annotation tools into the framework should be realized through the use of webservices (Alonso, 2004) using standard communication protocols.

Two critical and non-trivial issues for such a framework are annotation quality and the quality-control mechanisms. Unlike manually annotated corpora by domain experts, annotations by web users will likely be noisy. Although such annotated web corpora can still be utilized for manual curation, it would be desirable for the framework to provide analytical intelligence to make decisions about collating and resolving possibly conflicting and uncertain annotations from potentially numerous users and/or various NLP tools. This is an open area of research and deserves an active investigation.

2.2 Framework Architecture

The current implementation of the WebBANC framework consists of the following main components: an Annotation Server, the Annotation database, an OWL Ontology Interface, a Query and Retrieval Interface, and an Annotation Frontend.

The Annotation Frontend is a Firefox plug-in that uses XUL and JavaScript and supports two interfaces: one handles standard text and the other annotates PDF documents. The browser implementation allows distributed users to annotate websites. Users highlight words or phrases to annotate and link them to semantic tags by dragging or double-clicking the tag. The plain text interface builds upon the W3C Annotea project (Kahan et al., 2002). The PDF client leverages jPDFNotes (2008) and is compiled with Java 5.

The WebBANC framework lets developers expose any ontology by extending a Java class or implementing specific webservices. The OWL Ontology Interface sends available ontologies from the server to the Annotation Frontend through an OWL API. WebBANC uses OWL for ontology communication because it is a W3C standard and will allow others to develop new semantic tags and relationships as well as ease the development of new Annotation Frontends.

The Annotation Server handles communication between the Annotation Frontend and the backend database, which uses MySQL 5. Communication between clients and servers uses XML, and specifically either RDF or OWL, depending on the request context. The MySQL database is stored on an annotation server to support permanent storage and querying of manually annotated text. This allows NLP models to refine their prediction algorithms and also allows WebBANC to generate corpora in multiple formats. We intend to extend the framework with the ability to plug-in NLP models to support semi-automation, thereby allowing users to curate model-specific tags.

3 Results

We evaluated WebBANC at two levels: small-scale actual user annotation performance and large-scale simulation-based results. The purpose of the former is to determine the efficacy and accuracy of annotated corpora generated by untrained casual users. The latter was designed to draw conclusions regarding the diversity of user annotations generated on the Web and to compare the generated corpora with existing corpora in English, minority languages, and human generated categories, such as those found on Yahoo.com.

3.1 Small-Scale Study of Casual Annotators

To examine the effectiveness of untrained annotators using a web based annotation platform, WebBANC was released to several users. The purpose of this study was to test whether volunteer casual annotators are effective in terms of accuracy and throughput.

3.1.1 Evaluation Methodology

To examine the effectiveness of untrained annotators we conducted a study of users annotating web pages of their choosing for a named entity task. While annotating, users were restricted to the tags Person, Organization, and Location and were instructed to only use the system for fifteen minutes a day over four consecutive days. Users were also instructed for one of those days to annotate approximately 60 sentences extracted from the 2003 Conference on Natural Language Learning (CoNLL) training corpus with the same entity types; the sentences were un-tagged prior to the experiment. We refer to the training corpus as the CoNLL corpus, and selected it for our evaluation due to its widespread adoption as a benchmark corpus.

3.1.2 Small-Scale Study Results

The seven users created a corpus of 1,634 annotations: 1028 for general web pages and 606 for CoNLL data. Volunteer casual annotators with no previous annotation experience demonstrated high throughput, in comparison to the GENIA corpus (Kim et al., 2008).

Table 1: Recall and Precision for CoNLL annotations.

	Per	Loc	Org	Avg	CoNLL Avg
Recall (All Data)	1.00	0.94	0.82	0.92	0.81
Precision (All Data)	0.70	0.82	0.42	0.58	0.82

Table 2: Precision for CoNLL annotations with filtering.

	Per	Loc	Org	Avg
Precision (Majority Voting)	0.76	0.86	0.48	0.64
Precision (Coverage Req.)	0.73	0.90	0.55	0.69
Precision (Majority Voting + Coverage Req.)	0.79	0.95	0.69	0.79

While throughput is important, the accuracy of the annotations directly impacts the usefulness of the corpus. To test users' annotation accuracy we directly compared their annotations to the expertly created standard CoNLL corpus. Table 1 shows that the users collectively annotated every Person entity tagged by CoNLL, giving a recall of 1. User-level annotation of the Location entity also achieved a high recall of 0.94, but the Organization entity yielded a lower recall of 0.82. The average recall over the three entities is 0.92, which is an improvement over the average recall of 0.81 provided by the sixteen automated predictive tools in CoNLL.

User-level annotations demonstrated the following precision: 0.79, 0.95, and 0.69 for Person, Location, and Organization entities, respectively, with an average of 0.79. These results, shown in

Table 2, were calculated using majority voting after removing annotations with singular coverage. Based on users' feedback, annotating the Organization entity was the most unclear of the three. The average precision for the Person and Location entities was 0.87. Again, the casual user-level precision was comparable with the automated tools that attained an average precision of 0.82 over the three entities. For user-level annotations of arbitrary web pages of their choosing, 42.1%(31.2%) of the web pages were found the top 70(50) web pages viewed in the United States according to Alexa.com, an internet traffic rating site. Due to these results, the subsequent evaluation considered up to the top 100 websites in the United States in an effort to better represent possibly annotated websites. The webpage categories annotated included News, Politics, Technology, Blogs, Science, and others, showing a range of diverse entity types that casual users may annotate using WebBANC.

3.2 Large-Scale System Generated Simulations

Section 3.1 shows WebBANC's potential for high throughput and accuracy, but effectiveness is dependent on regularly visited web pages containing words that are useful to NLP annotated corpora. Therefore, our experiment compares the content of frequently visited web sites to established corpora.

3.2.1 Evaluation Methodology

For large-scale simulation-based evaluation, we conducted three experiments comparing different sets of corpora to web generated corpora. The first experiment identified human-curated categories using Yahoo.com, which has about twenty primary categories, such as Health, Politics, and Weather. The corpora generated from these categories allowed us to evaluate category-specific corpora, for example, a Sports corpus. The second experiment used the most commonly visited web sites for a minority language, specifically Icelandic, and compared the results to a half-million word Icelandic corpus published by the Institute of Lexicography in 1991 (Pind et al., 1991) and produced from the IFD, supplied by the Árni Magnússon Institute for Icelandic Studies. The final experiment is compared against the OANC to assess the potential for building general English corpora.

A simple examination of word counts and word diversity derived from web corpus annotations from popular websites can help determine the likelihood of creating a diverse corpus, and therefore assess whether the generated corpus is likely to be useful. However, the rate at which users collectively annotate words encountered during regular web browsing, which we call the *annotation percentage*, directly affects the expected word counts and will vary. We considered *annotation percentages* from 100%, 90%, ..., 50% to simulate different user scenarios.

The experiments contained simulations that permute the recursive depth searched, the annotation percentage, and the number *X* of frequently visited sites explored. To simulate the web pages a casual user might browse on a daily basis we used data from Alexa.com to identify the most popular X websites in the United States, where $X \in \{10, 25, 50, 100\}$, referred to as the *top X* sites. The depth is varied to simulate different user behavior; some users will only visit the main web page, while others will drill-down into sublevels. Corpora generated from depth 0 contain the text on the front page of each URL; depth 1 corpora contain all text from depth 0, and all text gathered by following URL links at depth 0; similarly, depth 2 corpora contain all text obtained from the depth 1 traversal, including text collected by following all links discovered at depth 1. The number of links, or URLs, harvested for the top 100 corpus at depth 2 (see Table 3) became too large to process and were left out of the results. We used the $wget$ Unix program to recursively follow these links. These 3 depths, 4 groupings of popularities and 6 annotation percentages generated 72 datasets (3*4*6=72) per corpus.

Table 3: The number of documents harvested for the top X corpora at each depth.

	Depth 0	Depth 1	Depth 2
Top 10	10	207	2,266
Top 25	25	940	16,576
Top 50	50	2,272	33,239
Top 100	100	5,047	111,188

We used *recall* and *precision* to compare performance in our top X generated corpora. Given an established corpus or category system, called a *base word list*, and a generated word list from the top X websites, we calculate precision and recall after the following pre-processing: for each site

in a URL list or base corpus, apply a Perl module from BootCaT (Baroni and Bernardini, 2004) to retrieve the text from that URL and remove all HTML tags; remove all punctuation and words that appear in the stop word filter; apply Porter stemming; and generate a unique term list.

3.2.2 Large-Scale Study Results

Human-Curated Corpora: To evaluate WebBANC's ability to generate category or entity specific corpora, we ran several simulations varying the traversal depth and quantity of top X sites. This experiment was designed to compare category recognition between the top X corpora and humanly-curated corpora. The results indicate that both the value of X and the traversal depth affect the quality of generated corpora.

Table 4 shows unique word counts for a select set of Yahoo.com categories including Nutrition, Sports, and Technology. Due to the limited text at depth 0 and the great expansion of text at depth 2, the decision was made to examine the human-curated categories at depth 1. There are fewer web pages to annotate at depth 1 than depth 2, and therefore depth 1 may better simulate likely user behavior.

Table 4: Unique word counts for human-curated corpora from Yahoo.com.

	Depth 0	Depth 1	Depth 2
Nutrition	417	7,071	16,452
Sports	796	17,760	74,840
Technology	432	16,440	100,163

As Table 4 shows, depth had less impact on the Nutrition corpus size. The pages retrieved at consecutive depths for Nutrition returned similar words, which negatively affected the uniqueness and diversity of the corpus. Sports benefited greatly from increased depth due to its hierarchical information content. Similar to Sports, information content for the Technology category was organized in a product-driven hierarchy, resulting in a higher dependence on the depth level.

We examined the top 100 most visited sites, compared them to three Yahoo.com human-curated corpora, both at depth 1, and examined the results with annotation percentages ranging from 100% to 50%. Figure 1 shows maximum recall of 67%, 70%, and 57% for the Nutrition, Sports, and Technology corpora, respectively. In the less

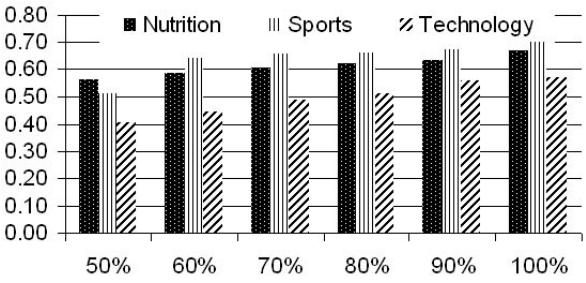

Figure 1: Recall of the top 100 corpus (depth 1) vs. human-curated Yahoo.com corpora (depth 1).

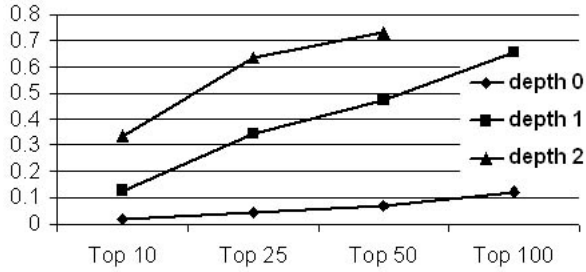

Figure 2: Recall of top X corpora at depths 0, 1, and 2 vs. Sports Yahoo.com corpus (depth 1) with annotation percentage of 70%.

ideal scenario, in which users collectively only annotate half of what they see, Figure 1 shows recall above 50% for two of the three Yahoo.com categories, indicating that users collectively annotating half of all encountered words can cover about half the possible words in a specialized corpus.

Finally, we examined recall of the top X corpora at different depths against the Sports corpus at depth 1 using an annotation percentage of 70% to demonstrate X's effect on word diversity. As Figure 2 shows, recall improved from depths 0 to 1 for the top 10 and top 100 sites by a factor of eight (1.6% to 12.9%) and a smaller factor of 5.3 (12.4% to 66.5%), respectively. The top 100 sites did not perform as well because increasing X decreases word uniqueness, which attenuates the benefits. As X increases for the top X sites, Figure 2 suggests that recall increases. The figure also shows that similar recall performance can be achieved at smaller X values by increasing the depth. fGiven that our results showed higher recall with larger X values and increased depth, it would be interesting to harvest larger numbers of websites in future work to determine if a saturation point for the number of documents examined exists.

Minority Language Corpora: The lack of an-

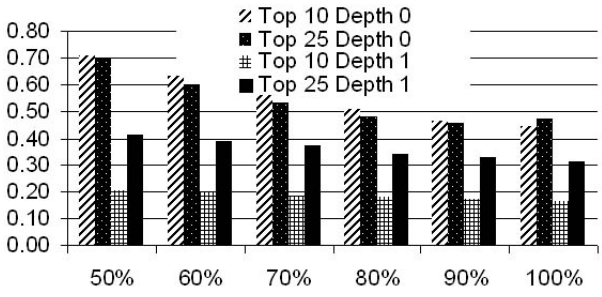

Figure 3: Precision of Icelandic top X corpora vs. IFD corpus.

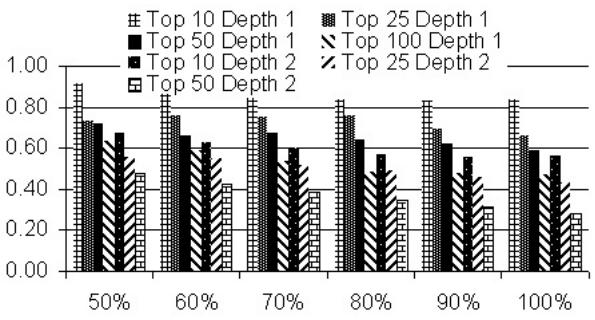

Figure 4: Precision comparison at different annotation percentages between OANC and the top X corpora.

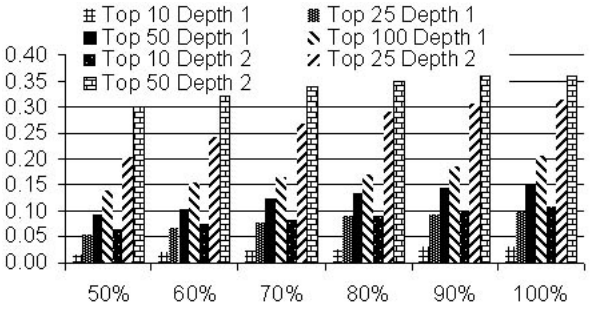

Figure 5: Recall comparison at different annotation percentages between OANC and the top X corpora.

notated corpora for minority languages is a primary cause for the dearth of machine learning tasks in these languages. The following experiment is designed to show that minority language speakers can annotate words during their daily browsing to aid in the construction of annotated corpora using the Icelandic language.

Figure 3 compares the precision of Icelandic top X corpora to the IFD corpus. The results suggest that words in the top X sites are useful for corpora generation, but diversity may be less than desirable, although 70% precision is attained for the top 10 and top 25 Icelandic websites at 50% annotation percentage. The results indicate that words encountered by Icelandic speakers in everyday web browsing may yield relatively precise Icelandic corpora.

Recall for Icelandic top X corpora is relatively low, around 30%, in comparison to the other experiments, for several reasons. Unlike the English corpora results, we did not apply a complete stemming or morphological tool, such as the Porter stemmer, and therefore many Icelandic words did not match their root words in the base corpora. In this simulation only a basic stemmer was applied (e.g. umlauts were not taken into account) causing some words to differ from their root words with the same semantic meaning. Future experiments on this topic should make use of newer lemmatization software for Icelandic, such as Lemmald (Ingason et al., 2008). IFD's use of literature is also a likely cause for the low recall since the most popular websites are news related.

The IFD corpus contained 35,883 unique words after applying a suffix stemmer and removing punctuation. Similar to the English corpora results for the top X sites, the Icelandic equivalents had low unique word counts for the top 10 (2,819) and top 25 (3,178) depth 0 searches, but increased at depth 1. For example, the top 25 contained 22,661 unique words, which more closely approximates the size of the IFD corpus. The majority of depth 0 corpora exclusively contain Icelandic words, however, examining corpora at depth 1 shows that other languages, mostly English, pollute the corpora due to depth 0 sites linking to web sites in other languages, although some English phrases are filtered by the stop word list.

General Corpora: To encourage corpus creation from the Web, it is important to determine if the Web represents the breadth of a particular language, which this experiment addresses by comparing the top X corpora to the OANC corpora.

Figure 4 suggests that the top X corpora may be useful, with precision values almost 70%, if users annotate text at the top X sites at depths 0 and 1. The precision values decline at depth 2; this may be caused by pages at increased depth containing more category specific language that does not represent American English as precisely.

The recall results in Figure 5 compare OANC, the base corpus, to the top X sites at depths 1 and 2 and show low performance, peaking at 36.2%.

This is partly caused by OANC being a balanced collection of texts, which includes categories seldom found in the top X sites, such as Fiction and Technical, although the results for X $\in \{25, 50\}$ at depth 2 represent dramatic improvements over depths 0 and 1.

The precision results support the hypothesis that the Web may be useful for annotating the American Nation Corpus (ANC) for specific genres or categories that are covered in-depth on the Web, such as Technology, Business, or Sports documents. However, the recall results validate work by Ide, Reppen and Suderman (Ide et al., 2002) claiming that general corpora constructed from web documents would not cover the same breadth of topics as the ANC, which is a testament to the scope of the ANC project.

4 Conclusion and Future Work

Annotated corpora generation presents a major obstacle to advancing modern Natural Language Processing technologies, especially for minority languages. In this paper we introduced the WebBANC framework, which aims to leverage a distributed web user community to build sufficiently diverse, semantically-rich, and large-scale corpora from user annotations. Accuracy and throughput were examined through a small-scale user study with promising results. We evaluated the diversity of the web-based corpora by comparing statistics against (a) corpora built from human-curated Yahoo.com categories, (b) a minority language corpus generated from the IFD, and (c) established domain corpora, such as OANC and CoNLL. Using up to 100 of the most commonly visited websites, according to Alexa.com, captured 35%-70% of the diversity of these base corpora at 70%-90% percent precision even using just half of the words encountered in these webpages. The actual user studies demonstrated a relatively high accuracy for the NER task that was comparable in performance to the majority of automatic NER tools.

The success of collaborate annotation projects, such as WebBANC rely heavily on user involvement. To increase the possibility of success for multi-lingual projects in the future we are developing other interfaces, such as collaborative games, that are beyond the scope of this paper. Collaborative annotation is likely to benefit from filtering and weighting techniques, as shown in Table 2, and our future work will incorporate inter-annotator agreement such as Kappa statistics.

Acknowledgments

We would like to thank the Fulbright Program for the year of support for Nathan Green in Iceland as well as the Árni Magnússon Institute for Icelandic Studies for supplying the Icelandic corpus. This work was funded by the Scientific Data Management Center under the Department of Energy's Scientific Discovery through Advanced Computing program. Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.

References

G. Alonso. Web Services: Concepts, Architectures and Applications. Springer, 2004.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25, 2000.

M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. *The 4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal, 2004.

S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference. W3C Recommendation, 10:2006–01, 2004.

L. Burnard. The Users Reference Guide for the British National Corpus. British National Corpus Consortium. *Oxford University Computing Service*, 1995.

S. DeRose, R. D. Jr., and E. Maler. XML Pointer Language (XPointer). *World Wide Web Consortium Working Draft*. March, 1998.

S. Helgadóttir. Testing data-driven learning algorithms for POS tagging of Icelandic. *Nordisk Sprogteknologi*, pages 2000–2004, 2004.

N. Ide and C. Macleod. The American National Corpus: A Standardized Resource of American English. *Corpus Linguistics*, pages 274–280, 2001.

N. Ide, R. Reppen, and K. Suderman. The American National Corpus: More than the web can provide. *The Third Language Resources and Evaluation Conference*, pages 839–844, 2002.

jPDFNotes Development Team. jPDFNotes. http://www.qoppa.com/pdfnotes/jpnindex.html.

J. Kahan, M. R. Koivunen, E. Prud'Hommeaux, and R. R. Swick. Annotea: an open RDF infrastructure for shared Web annotations. *Computer Networks*, 39:589–608, 2002.

J.D. Kim, T. Ohta, and J. Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10, 2008.

G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 10, 2004.

D. Krafzig, K. Banke, and D. Slama. Enterprise SOA: Service-Oriented Architecture Best Practices. *Prentice Hall Ptr*, 2004.

S. Krauwer. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. *The International Workshop Speech and Computer (SPECOM 200)*, Moscow, Russia, 2003.

Y. Labrou and T. Finin. Yahoo! as an ontology: using Yahoo! categories to describe documents. *The eighth international conference on Information and knowledge management*, pages 180–187, 1999.

F. Leitner and A. Valencia. A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Letters*, 582:1178–1181, 2008.

C. Mair. The corpus-based study of language change in progress: The extra value of tagged corpora. *The AAACL/ICAME Conference*, Ann Arbor, 2005.

J. Pind, F. Magnússon, and S. Briem. The Icelandic Frequency Dictionary. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland, 1991.

R. Ghani and D. Mladenic. Mining the Web to Create Minority Language Corpora. The 10th international conference on Information and knowledge management, pages 279–286, Athens, Georgia, 2001.

P. Rayson, J. Walkerdine, W. H. Fletcher, and A. Kilgarriff. Annotated web as corpus. A. Kilgarriff and M.Baroni, editors, *The 2nd International Workshop on Web as Corpus* (EACL06), pages 27–34, Trento, Italy, 2006.

E. Sang and F. D. Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *The seventh conference on Natural language learning at HLT-NAACL* 2003-Volume 4, pages 142–147, 2003.

K. Scannell. The Crúbadán Project: Corpus building for under-resourced languages. C. Fairon, H. Naets, A. Kilgarriff, and G.-M. de Schryver, editors, *Building and Exploring Web Corpora*, pages 5–15, Louvain-la-Neuve, Belgium, 2007.

V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4:14–28, 2006.

A. Ingason, S. Helgadóttir, H. Loftsson, and Eiríkur Rögnvaldsson A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). *Advances in Natural Language Processing*, 205–216, 2008.

T. Zesch, I. Gurevych, and M. Mühlhäuser. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. *Biannual Conference of the Society for Computational Linguistics and Language Technology*, pages 213–221, 2007.