The Shared Corpora Working Group Report

Adam Meyers
New York
University
New York, NY
meyers
at cs.nyu.edu

Nancy Ide
Vassar College
Poughkeepsie, NY
ide at cs.vassar.edu

Ludovic Denoyer
University of Paris
Paris, France
ludovic.denoyer
at lip6.fr

Yusuke Shinyama
New York
University
New York, NY
yusuke
at cs.nyu.edu

Abstract

We seek to identify a limited amount of representative corpora, suitable for annotation by the computational linguistics annotation community. Our hope is that a wide variety of annotation will be undertaken on the same corpora, which would facilitate: (1) the comparison of annotation schemes; (2) the merging of information represented by various annotation schemes; (3) the emergence of NLP systems that use information in multiple annotation schemes; and (4) the adoption of various types of best practice in corpus annotation. Such best practices would include: (a) clearer demarcation of phenomena being annotated; (b) the use of particular test corpora to determine whether a particular annotation task can feasibly achieve good agreement scores; (c) The use of underlying models for representing annotation content that facilitate merging, comparison, and analysis; and (d) To the extent possible, the use of common annotation categories or a mapping among categories for the same phenomenon used by different annotation groups.

This study will focus on the problem of identifying such corpora as well as the suitability of two candidate corpora: the Open portion of the American National Corpus (Ide and Macleod, 2001; Ide and Suderman, 2004) and the "Controversial" portions of the WikipediaXML corpus (Denoyer and Gallinari, 2006).

1 Introduction

This working group seeks to identify a limited amount of representative corpora, suitable for annotation by the computational linguistics annotation community. Our hope is that a wide variety of annotation will be undertaken on the same corpora, which would facilitate:

1. The comparison of annotation schemes
2. The merging of information represented by various annotation schemes
3. The emergence of NLP systems that use information in multiple annotation schemes; and
4. The adoption of various types of best practice in corpus annotation, including:
   (a) Clearer demarcation of the phenomena being annotated. Thus if predicate argument structure annotation adequately handles relative pronouns, a new project that is annotating coreference is less likely to include relative pronouns in their annotation; and
   (b) The use of particular test corpora to determine whether a particular annotation task can feasibly achieve good agreement scores.
   (c) The use of underlying models for representing annotation content that facilitate merging, comparison, and analysis.

Proceedings of the Linguistic Annotation Workshop, pages 184–190,
Prague, June 2007. ©2007 Association for Computational Linguistics

(d) To the extent possible, the use of common annotation categories or a mapping among categories for the same phenomenon used by different annotation groups.

In selecting shared corpora, we believe that the following issues must be taken into consideration:

1. The diversity of genres, lexical items and linguistic phenomena – this will ensure that the corpora will be useful to many different types of annotation efforts. Furthermore, systems using these corpora and annotation as data will be capable of handling larger and more varied corpora.

2. The availability of the same or similar corpora in a wide variety of languages;

3. The availability of corpora in a standard format that can be easily processed – there should be mechanisms in place to maintain the availability of corpora in this format in the future;

4. The ease in which the corpora can be obtained by anyone who wants to process or annotate them – corpora with free licenses or that are in the public domain are preferred

5. The degree with which the corpora is representative of text to be processed – this criterion can be met if the corpora is diverse (1 above) and/or if more corpora of the same kind is available for processing.

We have selected the following corpora for consideration:[1]

1. The OANC: the Open sections of the ANC corpus. These are the sections of the American National Corpus subject to the opened license, allowing them to be freely distributed. The full Open ANC (Version 2.0) contains about 14.5 megawords of American English and covers a variety of genres as indicated by the full pathnames taken from the ANC distribution (where a final 1 or 2 indicates which DVD the directory originates from):
   - spoken/telephone/switchboard
   - written_1/fiction/eggan
   - written_1/journal/slate
   - written_1/letters/icic
   - written_2/non-fiction/OUP
   - written_2/technical/biomed
   - written_2/travel_guides/berlitz1
   - written_2/travel_guides/berlitz2
   - written_1/journal/verbatim
   - spoken/face-to-face/charlotte
   - written_2/technical/911report
   - written_2/technical/plos
   - written_2/technical/government

2. The Controversial-Wikipedia-Corpus, a section of the Wikipedia XML corpus. WikipediaXML is a corpus derived from Wikipedia, converting Wikipedia into an XML corpus suitable for NLP processing. This corpus was selected from:
   - Those articles cited as controversial according to the November 28, 2006 version of the following Wikipedia page: http://en.wikipedia.org/wiki/Wikipedia: List_of_controversial_issues
   - The talk pages corresponding to these articles where Wikipedia users and the community debate aspects of articles. These debates may be about content or editorial considerations.
   - Articles in Japanese that are linked to the English pages (and the associated talk pages) are also part of our corpus.

2 American National Corpus

The American National Corpus (ANC) project (Ide and Macleod, 2001; Ide and Suderman, 2004) has released over 20 million words of spoken and written American English, available from the Linguistic Data Consortium. The ANC 2nd release consists of fiction, non-fiction, newspapers, technical reports, magazine and journal articles, a substantial amount of spoken data, data from blogs and other unedited web sources, travel guides, technical manuals, and other genres. All texts are annotated for sentence boundaries; token boundaries,

[1] These corpora can be downloaded from: http://nlp.cs.nyu.edu/wiki/corpuswg/SharedCorpora

lemma, and part of speech produced by two different taggers ; and noun and verb chunks. A subcorpus of 10 million words reflecting the genre distribution of the full ANC is currently being hand-validated for word and sentence boundaries, POS, and noun and verb chunks. For a complete description of the ANC 2nd release and its contents, see http://AmericanNationalCorpus.org.

Approximately 65 percent of the ANC data is distributed under an open license, which allows use and re-distribution of the data without restriction. The remainder of the corpus is distributed under a restricted license that disallows re-distribution or use of the data for commercial purposes for five years after its release date, unless the user is a member of the ANC Consortium. After five years, the data in the restricted portions of the corpus are covered by the open license.

ANC annotations are distributed as stand-off documents representing a set of graphs over the primary data, thus allowing for layering of annotations and inclusion of multiple annotations of the same type. Because most existing tools for corpus access and manipulation do not handle stand-off annotations, we have developed an easy-to-use tool and user interface to merge the user's choice of stand-off annotations with the primary data to form a single document in any of several XML and non-XML formats, which is distributed with the corpus. The ANC architecture and format is described fully in (Ide and Suderman, 2006).

2.1 The ULA Subcorpus

The Unified Linguistic Annotation (ULA) project has selected a 40,000 word subcorpus of the Open ANC for annotation with several different annotation schemes including: the Penn Treebank, PropBank, NomBank, the Penn Discourse Treebank, TimeML and Opinion Annotation.[2] This initial subcorpus can be broken down as follows:

- Spoken Language
  - charlotte: 5K words
  - switchboard: 5K words
- letters: 10K words
- Slate (Journal): 5K words
- Travel_guides: 5K words
- 911report: 5K words
- OUP books (Kaufman): 5K words

As the ULA project progresses, the participants intend to expand the corpora annotated to include a larger subsection of the OANC. They believe that the diversity of this corpus make it a reasonable testbed for tuning annotation schemes for diverse modalities. The Travel_guides and some of the slate articles have already been annotated by the FrameNet project. Thus the inclusion of these documents furthered the goal of producing a multiply annotated corpus by one additional project.

It is the recommendation of this working group that: (1) other groups annotate these same subcorpora; and (2) other groups choose additional corpora from the OANC to annotate and publicly announce which subsections they choose. We would be happy to put all such subsections on our website for download. The basic idea is to build up a consensus of what should be mutually annotated, in part, based on what groups choose to annotate and to try to get annotation projects to gravitate toward multiply annotated, freely available corpora.

3 The WikipediaXML Corpus

3.1 Why Wikipedia?

The Wikipedia corpus consists of articles in a wide range of topics written in different genres and mainly (a) *main* pages are encyclopedia style articles; and (b) *talk* pages are discussions about main pages they are linked to. The topics of these discussions range from editing contents to disagreements about content. Although Wikipedia texts are mostly limited to these two genres, we believe that it is well suited as training data for natural language processing because:

1. they are lexically diverse (e.g., providing a lot of lexical information for statistical systems);
2. the textual information is well structured
3. Wikipedia is a large and growing corpus

[2] Other corpora being annotated by the ULA project include sections of the Brown corpus and LDC parallel corpora.

4. the articles are multilingual (cf. section 3.4)
5. and the corpus has various other properties that many researchers feel would be interesting to exploit.

To date research in Computational Linguistics using Wikipedia includes: Automatic derivation of taxonomy information (Strube and Ponzetto, 2006; Suchanek et al., 2007; Zesch and Gurevych, 2007; Ponzetto, 2007); automatic recognition of pairs of similar sentences in two languages (Adafre and de Rijke, 2006); corpus mining (Rüdiger Gleim and Alexander Mehler and Matthias Dehmer, 2007), Named Entity Recognition (Toral and noz, 2007; Bunescu and Paşca, 2007) and relation extraction (Nguyen et al., 2007). In addition several shared tasks have been set up using Wikipedia as the target corpus including question answering (cf. (D. Ahn and V. Jijkoun and G. Mishne and K. Müller and M. de Rijke and S. Schlobach, 2004) and http://ilps.science.uva.nl/WiQA/); and information retrieval (Fuhr et al., 2006). Some other interesting properties of Wikipedia that have yet to be explored to our knowledge include: (1) Most main articles have talk pages which discuss them – perhaps this relation can be exploited by systems which try to detect discussions about topics, e.g., searches for discussions about current events topics; (2) There are various meta tags, many of which are not included in the WikipediaXML (see below), but nevertheless are retrievable from the original HTML files. Some of these may be useful for various applications. For example, the levels of disputability of the content of the main articles is annotated (cf. http://en.wikipedia.org/wiki/Wikipedia: Template_messages/Disputes).

3.2 Why WikipediaXML?

WikipediaXML (Denoyer and Gallinari, 2006) is an XML version of Wikipedia data, originally designed for Information Retrieval tasks such as INEX (Fuhr et al., 2006) and the XML Document Mining Challenge (Denoyer and P. Gallinari, 2006). WikipediaXML has become a standard machine readable form for Wikipedia, suitable for most Computational Linguistics purposes. It makes it easy to identify and read in the text portions of the document, removing or altering html and wiki code that is difficult to process in a standard way. The WikipediaXML standard has (so far) been used to process Wikipedia documents written in English, German, French, Dutch, Spanish, Chinese, Arabic and Japanese.

3.3 The Controversial Wikipedia Corpus

The English Wikipedia corpus is quite large (about 800K articles and growing). Frozen versions of the corpus are periodically available for download. We selected a 5 million word subcorpus which we believed would be good for a wide variety of annotation schemes. In particular, we chose articles listed as being controversial (in the English speaking world) according to the November 28, 2006 version of the following Wikipedia page: http://en.wikipedia.org/wiki/Wikipedia: List_of_controversial_issues. We believed that controversial articles would be more likely than randomly selected articles to: (1) include interesting discourse phenomena and emotive language; and (2) have interesting "talk" pages (indeed, some of Wikipedia pages have no associated talk pages).

3.4 The Multi-linguality of Wikipedia

One of the main good points of Wikipedia is the fact that it is a very large multilingual resource. This provides several advantages over single-language corpora, perhaps the clearest such advantage being the availability of same-genre/same-format text for many languages. Although, Wikipedia in languages other than English do not approach 800K articles in size, there are currently at least 14 languages with over 100K entries.

It should be clear however, that it is definitely not a parallel corpus. Although pages are sometimes translated in their entirety, this is the exception, not the rule. Pages can be partially translated or summarized into the target language. Individually written pages can be linked after they are created if it is believed that they are about the same topic. Also, initially parallel pages can be edited in both languages, causing them to diverge. We therefore decided to do a small small pilot study to attempt to characterize the degree of similarity between English articles in Wikipedia and articles written in other languages that have been linked. There are 476 English Wikipedia articles in the Controversial corpus

Classification	Frequency
Totally Different	2
Same General Topic	3
Overlapping Topics	11
Same Topics	33
Parallel	1

and 384 associated "talk" pages. There are approximately 10,000 articles of various languages that are linked to the English articles. We asked some English/Japanese bilingual speakers to evaluate the degree of similarity of as many of the the 305 Japanese articles that were linked to English controversial articles. As of this date, 50 articles were evaluated with the results summarized as table 3.4.[3] These preliminary results suggest the following:

- Languate-linked Wikipedia would usually be classified as "comparable" corpora as 34 (68%) of the articles were classified as covering the same topics or being parallel.
- It may be possible to extract a parallel corpus for a given pair of languages from Wikipedia. If the above sample is representative, approximately 2% of the articles are parallel. (While the existance of one parallel article does not provide statistically significant evidence that 2% of Wikipedia is parallel, the article's existance is still significant.) Furthermore, additional parallel sentences may be extracted from some of the other comparable articles using techniques along the lines of (Adafre and de Rijke, 2006).

Obviously, a more detailed study would be necessary to gain a more complete understanding of how language-linked articles are related in Wikipedia.[4] Such a study would include characterizations of all linked articles for several languages. This study could lead to some practical applications, e.g., (1) the creation of parallel subcorpora for a number of languages; (2) the selection of an English monolingual subcorpus consisting of articles, each of which is parallel to some article in some other language; etc.; (3) A compilation of parallel sentences extracted from comparable articles. While parallel subcorpora are of maximal utility, finding parallel sentences could still be extremely useful. (Adafre and de Rijke, 2006) reports one attempt to automatically select parallel Dutch/English sentences from language-linked Wikipedia articles with an accuracy of approximately 45%. Even if higher accuracy cannot be achieved, this still suggests that it is possible to create a parallel corpus (of isolated sentences) using a combination of automatic and manual means. A human translator would have to go through proposed parallel sentences and eliminate about one half of them, but would not have to do any manual translation. Selection of corpora for annotation purposes depends on a number of factors including: the type of annotation (e.g., a corpus of isolated sentences would not be appropriate for discourse annotation); and possibly an application the annotation is tuned for (e.g., Machine Translation, Information Extraction, etc.)

It should be noted that the corpus was chosen for the controversialness of its articles in the English-speaking community. It should, however, not be expected that the same articles will be controversial in other languages. More generally, the language-linked Wikipedia articles may have different cultural contexts depending on the language they are written in. This is an additional feature that we could test in a wider study. Furthermore, English pages are somewhat special because they're considered as the common platform and expected to be neutral to any country. But other lanauages somewhat reflects the view of each country where the language is spoken. Indeed, some EN articles are labeled as *USA-centric* (cf. http://en.wikipedia.org/wiki/Category:USA-centric).

Finally, our choice of a corpus based on controversy may have not been the most efficient choice if our goal had been specifically to find parallel corpora. Just as choosing corpora of articles that are controversial (in the English-speaking world) may have helped finding articles interesting to annotate it is possible that some other choice, e.g., technical articles, may have helped select articles likely

[3] According to www.wikipedia.org there are currently over 350K Japanese articles.

[4] Long Wikipedia articles may be split into multiple articles. This can result in N to 1, or even N to N, matches between language-linked articles if a topic is split in one language, but not in another.

to be translated in full[5] Thus further study may be required to choose the right Wikipedia balance for a set of priorities agreed upon by the annotation community.

4 Legal Issues

The American National Corpus has taken great pains to establish that the open subset of the corpus is freely usable by the community. The open license[6] makes it clear that these corpora can be used for any reason and are freely distributable.

In contrast, some aspects of the licensing agreement of corpora derived from Wikipedia are unclear. Wikipedia is governed by the GNU Free Document License which includes a provision that "derived works" are subject to this license as well. While most academic researchers would be uneffected by this provision, the effect of this provision is unclear with respect to commercial products.

Under one view, a machine translation system that uses a statistical model trained on Wikipedia corpora is not derived from these corpora. However, on another view it is derived. We contacted Wikipedia staff by letter asking for clarification on this issue and received the following response from Michelle Kinney on behalf of Wikipedia information team:

> Wikipedia does not offer legal advice, and therefore cannot help you decide how the GNU Free Documentation License (GFDL) or any other free license applies to your particular situation. Please contact a local bar association, law society or similar association of jurists in your legal jurisdiction to obtain a referral to a competent legal professional.
>
> You may also wish to review the full text of the GFDL yourself:
>
> http://en.wikipedia.org/wiki/Wikipedia:
> Text_of_the_GNU_Free_Documentation_License

While some candidate corpora are completely in the public domain, e.g., political speeches and very old documents, many candidate corpora are under the GFDL or similar "copyleft" licenses. These include other licenses by the GNU organization and several Creative Commons licenses. It is simply unclear how copyleft licenses should be applied to corpora used as data in computational linguistics and we believe that this is an important legal question for the Computational Linguistics community. In addition to Wikipedia, this issue effects a wide variety of corpora (e.g., other wiki corpora, some of the corpora being developed by the American National Corpus, etc.).

However, getting such legal opinions is expensive and has to be done carefully. Hypothetically, suppose NYU's legal department wrote an opinion letter stating that products that were not corpora themselves were not to be considered derived works for purposes of some list of copyleft licensing agreements. Furthermore, let's suppose that several annotation projects relied on this opinion and produced millions of dollars worth of annotation for one such corpus. Large corporations still might not use these corpora unless their own legal departments agreed with NYU's opinion. For the annotation community, this could mean that certain annotation would only be used by academics and not by industry, and most annotation researchers would not be happy with this outcome. It therefore may be worth some effort on the part of whole NLP community to seek some clear determinations on this issue.

5 Concluding Remarks

The working group selected two freely distributable corpora for purposes of annotation. Our goal was to choose texts for annotation by multiple annotation research groups and describe the process and the pitfalls involved in selecting those texts. We, furthermore, aimed to establish a protocol for sharing texts, so that the same texts are annotated with multiple annotation schemes. This protocol cannot be setup carte blanche by this group of researchers. Rather, we believe that our report in combination with the discussion at the upcoming meeting of the Lingustic Annotation Workshop will provide the jumpstart necessary for such a protocol to be put in place.

[5] Informally, we observe that linked Japanese/English pairs of articles about abstract topics (e.g., Adultery, Agnosticsism, Antisemitism, Capitalism, Censorship, Catholicism) are less likely to contain parallel sentences than articles about specific events or people (e.g., Adolf Hitler, Barbara Streisand, The Los Angeles Riots, etc.)

[6] http://projects.ldc.upenn.edu/ANC/ANC_SecondRelease_EndUserLicense_Open.htm

References

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. In *EACL 2006 Workshop: Wikis and blogs and other dynamic text source*, Trento, Italy.

Razvan Bunescu and Marius Pasça. 2007. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proc. of NAACL/HLT 2007*.

D. Ahn and V. Jijkoun and G. Mishne and K. Müller and M. de Rijke and S. Schlobach. 2004. Using Wikipedia at the TREC QA Track. In *Proc. TREC 2004*.

Ludovic Denoyer and Patrick Gallinari. 2006. The Wikipedia XML Corpus. *SIGIR Forum*.

L. Denoyer and A. Vercoustre P. Gallinari. 2006. Report on the XML Mining Track at INEX 2005 and INEX 2006 : Categorization and Clustering of XML Documents. In *Advances in XML Information Retrieval and Evaluation: Fifthth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX'06)*.

N. Fuhr, M. Lalmas, and S. Malik. 2006. Advances in XML Information Retrieval and Evaluation. In *5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*.

N. Ide and C. Macleod. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, Lancaster, UK.

N. Ide and K. Suderman. 2004. The american national corpus first release. In *Proceedings of LREC 2004*, pages 1681–1684, Lisbon, Portugal.

N. Ide and K. Suderman. 2006. Integrating linguistic resources: The american national corpus model. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Genoa, Italy.

D. P.T. Nguyen, Y. Matsuo, and M. Ishizuka. 2007. Subtree Mining for Relation Extraction from Wikipedia. In *Proc. of NAACL/HLT 2007*.

Simone Paolo Ponzetto. 2007. Creating a Knowledge Base From a Collaboratively Generated Encyclopedia. In *Proc. of NAACL/HLT 2007*.

Rüdiger Gleim and Alexander Mehler and Matthias Dehmer. 2007. Web Corpus Mining by instance of Wikipedia. In *Proc. 2nd Web as Corpus Workshop at EACL 2006*.

M. Strube and S. P. Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*, pages 1419–1424.

F. M. Suchanek, G. Kasneci, and G.Weikum. 2007. YAGO: A core of semantic knowledge. In *Proc. of WWW-07*.

Antonio Toral and Rafael Mu noz. 2007. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proc. of NAACL/HLT 2007*.

Torsten Zesch and Iryna Gurevych. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proc of NAACL-HLT 2007 Workshop: TextGraphs-2*.