Increasing Interoperability for Embedding Corpus Annotation Pipelines in Wmatrix and other corpus retrieval tools

Paul Rayson
School of Computing and Communications, Lancaster University, Lancaster, UK
p.rayson@lancaster.ac.uk

Abstract

Computational tools and methods employed in corpus linguistics are split into three main types: compilation, annotation and retrieval. These mirror and support the usual corpus linguistics methodology of corpus collection, manual and/or automatic tagging, followed by query and analysis. Typically, corpus software to support retrieval implements some or all of the five major methods in corpus linguistics only at the word level: frequency list, concordance, keyword, collocation and n-gram, and such software may or may not provide support for text which has already been tagged, for example at the part-of-speech (POS) level. Wmatrix is currently one of the few retrieval tools which have annotation tools built in. However, annotation in Wmatrix is currently limited to the UCREL English POS and semantic tagging pipeline. In this paper, we describe an approach to extend support for embedding other tagging pipelines and tools in Wmatrix via the use of APIs, and describe how such an approach is also applicable to other retrieval tools, potentially enabling support for tagged data.

Keywords: corpus annotation, tagging, API

1. Introduction

Many different computational tools are used to support corpus linguistics research. Typically, these fit into one of three main categories. First, *compilation* tools are used to support corpus collection, and these include transcription, OCR, scanning and encoding tools. In the web-as-corpus paradigm, this category also includes web scraping and cleaning tools to collect data from web pages, online forums or social media along with tools to remove boilerplate or duplicated text, e.g. WebBootCaT (Baroni et al., 2006). Second, once a corpus is compiled, it may need to be *annotated* at one or more levels for later linguistic analysis. A common such level is part-of-speech (POS) annotation which has proved fruitful over the years for grammatical analysis and as a stepping stone to other higher levels such as semantic tagging. Annotation may be applied manually by one person or collaboratively by a team (e.g. using such tools as eMargin[1] or Brat[2]), and/or automatically using pre-existing tagging software (e.g. CLAWS (Garside and Smith, 1997)). The final category of corpus software is the most often used and cited in corpus papers (see Rayson (2015) for a quantitative in-depth survey), that of corpus *retrieval*. Corpus retrieval software began life[3] around 50 years ago with computerised concordances in key-word-in-context (KWIC) format, and steadily gained extra features such as frequency lists, keywords, collocations and n-grams. Recent developments, notably demonstrated by many papers at the Corpus Linguistics 2017 conference in Birmingham, have been to bring in tools and methods from other areas such as Natural Language Processing, such as topic modelling, or for researchers to develop their own software, or use other scripting languages (such as Python or R) to carry out analyses (as pioneered by Baayen (2008) and Gries (2013)). In this paper, we restrict ourselves to the pre-existing and widely available corpus query engines and retrieval tools.

2. Limitations of existing retrieval tools

One important limitation with many corpus retrieval tools is their ability to deal only with raw unannotated corpora and provide results only at the word level. This reduces the power of queries to surface patterns in the text and fails to take advantages of lemma searches which depends on POS analysis to link surface forms to dictionary headwords. Rayson (2008) and Culpeper (2009) have also shown the advantages of performing keyness analysis beyond the level of the word by combining the key words approach pioneered by Scott (1997) with semantic annotation. Workarounds with regular expressions do permit some of the existing desktop corpus query tools (such as WordSmith and AntConc) to work with tagged data, but it is the web-based corpus retrieval systems such as BYU (Davies, 2005) and CQPweb (Hardie, 2012) which have sufficient storage, power and complexity to more fully exploit tagged corpora.

The second major restriction, even with some existing web-based retrieval systems, is that corpus data must be tagged before it can be loaded (if the tool supports upload of new data directly) in or indexed by these tools. Only a few tools combine corpus annotation tools with corpus retrieval methods, and for ease of use by non-technical users, this combination offers many advantages and a shallow learning curve. Sketch Engine[4] incorporates POS taggers and lemmatisers for many languages and text is automatically tagged after upload or during processing of web-derived corpora. LancsBox (Brezina et al., 2015) version 3 also now incorporates the TreeTagger[5] in order to POS tag and lemmatise some languages. Wmatrix, through its Tag Wiz-

[1] https://emargin.bcu.ac.uk/

[2] http://brat.nlplab.org/

[3] See http://timemapper.okfnlabs.org/muranava/history-of-computerised-corpus-tools for an excellent visual time line produced by Mura Nava.

[4] https://www.sketchengine.co.uk/

[5] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

ard feature permits the upload and automatic POS and semantic tagging of English corpora.

3. Improving Interoperability

Corpus linguists have noted that different tools produce different results e.g. even in terms of calculating the size of a corpus (Brezina and Timperley, 2017) due to tokenisation differences resulting from a programmer's decisions encoded in software or by corpus compilers methods. For reproducibility and replication purposes there are many advantages to be gained from comparing and retaining separate implementations of standard corpus methods. However, corpus software development has now reached a level of maturity where good software development practices of design, implementation, distribution and component reuse should be adopted.

Some previous research into interoperability of corpus methods has been limited and small scale, and focussed on potential quick wins for linking analysis components in a small group of web-based tools (Wmatrix, CQPweb, IntelliText, and WordTree) (Moreton et al., 2012). By connecting such tools together, we are not just improving interoperability and reusability, but this will enable researchers to try out research methods and tools that are established in other disciplinary communities but are not so familiar in their own. For example, there are many similar tools to those developed in corpus linguistics which have long been employed in other areas that are not so well known to corpus researchers, from at least three other areas: (a) Computer Assisted Qualitative Data Analysis (CAQDAS) tools such as ATLAS.ti, Nvivo, Wordstat (b) Psycholinguistics software such as Linguistic Inquiry and Word Count (LIWC) and (c) Digital Humanities tools such as Voyant and MONK.

Amongst other work on interoperability is an ongoing effort to develop a Corpus Query Lingua Franca (CQLF) ISO standard[6] for corpus query formatting although this is not adopted in any query tools, and Vidler and Wattam (2017) have proposed a metadata standard for describing properties of corpus files and resources to enable better sharing and reuse of data between tools. In the remainder of the paper, we focus on addressing limitations related to the lack of flexible annotation facilities in corpus query tools. Other options for interoperability and reproducibility would be to share retrieval method modules directly.

4. Linking corpus annotation pipelines to retrieval tools

In order to improve interoperability between tagging pipelines and retrieval tools, we propose the use of Application programming interfaces (APIs)[7]. An API can be created to send raw corpus data to a remote server, where it is tagged, and then return the result. Such an approach will enable support for taggers and tagging pipelines to be incorporated not only into web-based corpus retrieval tools, but also in downloadable desktop applications. Web-based software is hosted on more powerful servers where taggers could more easily be housed alongside retrieval systems but for downloadable software such as AntConc and WordSmith, a user's personal computer may not be powerful enough to run large tagging processes in reasonable amounts of time. Even with web-based systems on remote servers, there may be a requirement to run extremely large tagging jobs in other parallel systems such as high performance clusters or Hadoop/SPARK Map Reduce frameworks (Wattam et al., 2014). This is where an API-based approach has important advantages since it enables the separation of the corpus processing and is independent of platform so existing Linux-based taggers can be linked for users running Windows locally, for example.

APIs have already been embedded in a small number of corpus tools to link with the compilation phase rather than the annotation phase. For example, in Laurence Anthony's AntCorGen tool[8], he has implemented an API link to the PLOS One research database to enable searching and download of academic papers which can be turned into a corpus. Also, Laurence's FireAnt[9] employs the Twitter Streaming API similarly to download Tweets and collate them into a corpus for later analysis.

4.1. Wmatrix case study

Wmatrix[10] is a corpus software tool combining annotation and retrieval methods. It provides a web interface to the existing USAS semantic tagger and CLAWS part-of-speech tagger corpus annotation tools, and standard corpus linguistic methodologies such as frequency lists, key words and concordances. It also extends the keywords method to key grammatical categories and key semantic domains by combining the taggers with the keyness metric (Rayson, 2008). Wmatrix allows the non-technical user to run these tools in a tag wizard via a web browser such as Chrome, and so will run on any computer (Mac, Windows, Linux, Unix) with a network connection. Earlier versions were available for Unix via terminal-based command line access (tmatrix) and Unix via X-windows (Xmatrix), but these only offered retrieval of text pre-annotated with USAS and CLAWS. With the incorporation of the two taggers, Wmatrix was designed for the analysis of English data only and has been used for a wide range of projects from learner data, interview transcript analysis, fiction and non-fiction corpora.

Wmatrix includes multiple components written in different programming languages, see Figure 1 for the architecture diagram of the current system. The two taggers, CLAWS and USAS, are written in C. The frequency profiling, concordancing and keyness comparison tools are also written in C. The collocation tool is developed in Java. Unix shell scripts and Perl scripts act as glue to link all these components together to the web front end. Underlying the system, the corpus data is stored in a Unix file system. User access is currently controlled using Apache's basic authentication (htaccess). The current version, Wmatrix3, is hosted on Lancaster University's cloud infrastructure on a Debian virtual machine.

[6] https://www.iso.org/standard/37337.html?browse=tc

[7] https://en.wikipedia.org/wiki/Application_programming_interface

[8] http://www.laurenceanthony.net/software/antcorgen/

[9] http://www.laurenceanthony.net/software/fireant/

[10] http://ucrel.lancs.ac.uk/wmatrix/

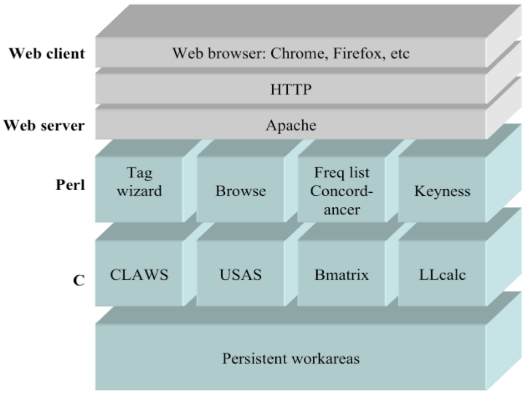

Figure 1: Wmatrix architecture.

A key and imminent requirement is to allow Wmatrix to be used with languages other than English. Semantic taggers already exist for a number of other languages (Russian and Finnish) and in recent years, we have been bootstrapping semantic taggers for an increasing number of languages (Piao et al., 2015; Piao et al., 2016; Piao et al., 2017a) and historical time periods (Piao et al., 2017b). Linguistic resources for these semantic taggers are freely available on GitHub[11] and many of them have been made available via a SOAP API and a downloadable GUI[12]. Our proposal is to create REST APIs for these and other tools[13] and add a new module into the existing Wmatrix architecture which sends data to be tagged via these APIs. This would sit alongside the existing tag wizard and remove the need to install other taggers directly on the Wmatrix server alongside the CLAWS and USAS modules. Non-English data which cannot be tagged through the existing tag wizard will then be directed to these components sitting on other servers for tagging. If we retain the two levels of annotation (POS and semantic tagging) then the existing infrastructure of Wmatrix will be sufficient, however further levels of annotation (e.g. dependency parsing) will require additional database or indexing operations to be implemented.

Once this architecture is in place, sample code will be made available on the API server to allow other corpus retrieval tools to access the taggers. Similar APIs could be used to permit other pipelines such as Stanford Core, GATE and OpenNLP to be linked. Further development of desktop tools would be needed for them to become annotation-aware once they can tag data via the APIs. In terms of feasibility of implementation of the REST APIs and incorporation into Wmatrix, this is fairly low risk. We have been running and supporting the CLAWS web based tagger since 1998, and the SOAP APIs for the multilingual USAS taggers for three years, and APIs have become widely adopted for NLP analytics tools and websites.

5. Conclusion

In this paper, we have outlined a proposal for enabling interoperability between two major types of corpus linguistics software, the annotation and retrieval (or query) systems. Currently, there are only a handful of tools which incorporate some form of both types, and previous research has shown the benefits of corpus queries operating beyond the level of the word. To improve ease of use for non-technical users, we propose the embedding and linking of further corpus annotation pipelines so that end-users can add annotation to their own data and then continue to use their current preferred tools for research. This proposed approach will contribute to improved interoperability in the corpus software community by simplifying the addition of new methods to existing tools and is complementary to other efforts to foster exchangeable and reusable components across corpus platforms. Other potential options, such as installing the taggers locally (rather than Wmatrix itself), can be explored, but it is expected that the API route would be preferable to tool developers and end-users alike.

6. Acknowledgements

Wmatrix was initially developed within the REVERE project (REVerse Engineering of Requirements) funded by the EPSRC, project number GR/MO4846, 1998-2001. Ongoing maintenance of taggers (e.g. Linux porting work by Stephen Wattam), development of new components (e.g. L-gram developed by Eddie Bell and C-grams developed by Andrew Stone) and dictionary updates (e.g. by Sheryl Prentice) are funded by user licence fees. Semantic taggers for new languages are being developed by Scott Piao and funded by the UCREL research centre, in collaboration with a large number of researchers worldwide[14]. Metaphor extensions have been developed in the MELC project (Metaphor in end-of-life care) funded by the ESRC (grant reference ES/J007927/1). The Historical Thesaurus Semantic Tagger (HTST) was developed in the SAMUELS project (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) funded by the AHRC in conjunction with the ESRC (grant reference AH/L010062/1). The Welsh Semantic Tagger research is funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes; The National Corpus of Contemporary Welsh) Project (Grant Number ES/M011348/1).

7. Bibliographical References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceedings of Euralex*, Turin.

Brezina, V. and Timperley, M. (2017). How large is the BNC? a proposal for standardised tokenization and word counting. In *Proceedings of Corpus Linguistics 2017*, Birmingham, UK.

[11] https://github.com/UCREL/Multilingual-USAS

[12] http://ucrel.lancs.ac.uk/usas/

[13] to be hosted on http://ucrel-api.lancs.ac.uk/

[14] See http://ucrel.lancs.ac.uk/usas/ for details.

Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Culpeper, J. (2009). Keyness: words, parts-of-speech and semantic categories in the character-talk of shakespeare's romeo and juliet. *International Journal of Corpus Linguistics*, 14(1):29–59.

Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, 10(3):307–334.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Roger Garside, et al., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121. Longman, London.

Gries, S. T. (2013). *Statistics for linguistics with R*. Mouton De Gruyter, Berlin and New York, 2nd edition.

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Moreton, E., Nesi, H., Rayson, P., Sharoff, S., and Stephenson, P. (2012). Corpus tools interoperability survey. Technical report.

Piao, S., Bianchi, F., Dayrell, C., D'egidio, A., and Rayson, P. (2015). Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1268–1274. Association for Computational Linguistics, 6.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, R., Shafi, J., Teh, P., and Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*, pages 2614–2619. European Language Resources Association (ELRA).

Piao, S., Rayson, P., Knight, D., Watkins, G., and Donnelly, K. (2017a). Towards a Welsh semantic tagger: creating lexicons for a resource poor language. In *Proceedings of Corpus Linguistics 2017*.

Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., and Alexander, M. (2017b). A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language*, 46:113–135.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Rayson, P. (2015). Computational tools and methods for corpus compilation and analysis. In Douglas Biber et al., editors, *The Cambridge Handbook of English corpus linguistics*, pages 32–49. Cambridge University Press.

Scott, M. (1997). PC analysis of key words - and key key words. *System*, 25(2):233–245.

Vidler, J. and Wattam, S. (2017). Keeping properties with the data: CL-MetaHeaders - an open specification. In Piotr Bański, et al., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017*, pages 35–41.

Wattam, S., Rayson, P., Alexander, M., and Anderson, J. (2014). Experiences with parallelisation of an existing nlp pipeline: tagging hansard. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*, pages 4093–4096. ELRA.