Linguistic Resource Creation in a Web 2.0 World

Jordan Boyd-Graber*

jbg@umiacs.umd.edu

University of Maryland

iSchool and Institute for Advanced Computer Studies

WordNet [Miller, 1990] remains an important resource for natural language processing [Kilgarriff, 2000], but it is changing. Not in its structure, but in how its represented. For over a decade, WordNet has been stored in flat text files that have been accessed by special purpose tools. Now, WordNet is moving to a relational database format. This transition offers a number of opportunities to bring the curation and development into the Web 2.0 world. If done correctly, this transition would allow WordNet to be a resource that grows based on user feedback, the input and insights of researchers, and the synergies between the two.

WordNet is important to me, and I want to see it succeed. I've used it in my research [Boyd-Graber et al., 2007, Boyd-Graber and Resnik, 2010], I've been active interacting with it in the context of the open source community, I've used it in my teaching, and I've seen, during my time at Princeton as a graduate student, the evolution and development of WordNet. I'll limit myself here to WordNet for the sake of concreteness; however, I'm sure that many of the examples I mention here are applicable to other resources, linguistic or otherwise.

*Jordan Boyd-Graber is supported by the Army Research Laboratory through ARL Cooperative Agreement W911NF-09-2-0072 and by NSF grant #1018625. Any opinions, findings, conclusions, or recommendations expressed are the author's and do not necessarily reflect those of the sponsors.

1 Decentralized Infrastructure Development

As the data that serve as the foundation of WordNet change, its programmatic interface also has to change. Traditionally, the programmatic interface has been one designed for an end user who wants to treat WordNet as a dictionary. The explosion of research on WordNet revealed thousands of other uses, but the official API has not grown to accommodate these novel uses. As a result, countless ad hoc interfaces have been reinvented over the years.

This is because continued growth and development of WordNet is limited by the cyclical focus of funding agencies. It is also limited to the closed computing environment of Princeton University. Many open source projects have successfully blossomed from academic endeavors, and there is a strong infrastructure in place for coordinating volunteer (and paid) contributions from around the globe. Moving WordNet development to a platform like Sourceforge or Google Code would increase visibility to the non-academic community and would allow for greater involvement in efforts like Google's summer of code.

By making sure the overhaul uses such a framework for its development process, future modifications will be less painful and better able to leverage the open source community. For instance, I am one of the maintainers of the Natural Language Toolkit's [Loper and Bird, 2002] interface to WordNet, which is developed in a decentralized environment that welcomes outside enhancements. As a result, NLTK has encouraged WordNet's integration into undergraduate education, hobbyist projects, and online projects.

2 User Additions

There has been plenty of hype about "Web 2.0," and not every project can benefit from crowd sourcing. However, projects like Wikipedia have shown that distilling real-world knowledge into a systematic form can definitely benefit from an interactive web environment.

Even with the significant overhead of composing and sending an e-mail, thousands have offered advice and criticism on the contents of WordNet, at times overwhelming the maintainers. Currently, changes only appear in new versions of WordNet. This is important for scientific reproducibility, and there would need

to continue to be reference versions (snapshots) to compare against. However, a public-facing, mutable WordNet would show that it's a living resource and encourage high throughput of changes suggested by users.

WordNet Benefits from User Additions The following means of improving WordNet could be offered by a more interactive interface to WordNet:

1. Report words missing form a synset
2. Report links missing between synsets and senses
3. Suggest missing synsets
4. Linking WN senses to other relevant knowledge sources: Wiktionary (in multiple languages), Wikipedia, etc.
5. Giving examples of usage

For each of these inputs, allowing users to vote on the quality of suggested changes would allow the cream to float to the top and ease the efforts of those maintaining the data in WordNet.

Points 1-3 are important, but they don't revolutionize WordNet in any substantial way. Points 4-5 do. They create larger sense annotated corpora. Building point 5 is particularly compelling to create diverse, contemporary sense annotated corpora. WordNet could use the following model to solicit such annotations (inspired by the resource LabelMe [Russell et al., 2008]):

- When a user downloads WordNet, ask them to specify an e-mail address (this also gives the added benefit of collecting statistics about downloaders)
- Send a link to a download in 30 minutes
- Or, give them the option of finding five contexts where a randomly selected sense appears on the web. If they do this, give them the download immediately

Alternative Arguments for Adopting an Interactive Web Environment Even if nobody does this voluntarily (which I find hard to believe), creating this infrastructure in the overhaul of WordNet will still offer benefits. For example, suppose that someone wants to create a sense tagged corpus. If they're using WordNet as the underlying sense inventory, WordNet's web interface will exist, which they

can then bootstrap (e.g. using a crowd sourcing platform like Mechanical Turk) and then get their results in the next released version of WordNet.

Such an outcome is a win for everyone; the person building the corpus gets a cheaper development process and WordNet and WordNet's many users get additional data that would otherwise be left to gather dust somewhere.

Overhead and Maliciousness Of course, there are many subtleties in this setup that I am glossing over. One wants to ensure a level of data purity, which would require review by either professional maintainers or by deputized editors (as is done in Wikipedia). This would require additional procedures, additional interface considerations, etc. These challenges, however, are not insurmountable, and I think are worth the possible rewards.

One thing to keep in mind is that the goal of this is to build a community that will police itself. There will be jokers, vandals, and those seeking to game the system. These will, with appropriate standards and guidance as seen on other collaborative websites, be outweighed by thoughtful contributors who want to see the resource evolve and mature.

3 Academic and Industrial Contributions

WordNet has often been used as a building block. It has been translated into many languages [Hamp and Feldweg, 1997, Ordan and Wintner, 2007, Sagot and Fišer, 2008] and extended [Snow et al., 2006, Denecke, 2008, Boyd-Graber et al., 2006] to include news aspects and domains. As WordNet grows, it should also be designed to accommodate existing and future extensions as a part of the WordNet ecosystem.

As an example of how cumbersome it presently is to extend WordNet, instructions for using the Stanford WordNet project involve downloading the official WordNet and overwriting a directory. In contrast, using a relational database allows for all of the myriad projects that extend or align WordNet to be distributed along with WordNet. A project could define new edges, new links, or annotations to WordNet. When a user chooses to download WordNet, he or she could choose to download either the "base" WordNet only or also get an additional table(s) that contains selected third-party additions.

The integration of these third-party additions would be facilitated by the distributed infrastructure development suggested in Section 1, which would be made available to authorized developers. Moreover, the WordNet webpage could be a clearinghouse for previewing and learning about these extensions. If the expectation for any system built on WordNet would be that it would eventually reside in this open, free ecosystem, extensions and translations of WordNet would be more available, accessible, and usable by the community than they are today.

Conclusion

The value of linguistic resources is that they are created, curated, and refined by experts who are able to create theoretically sound repositories of knowledge. This is a noble service that creates invaluable resources. When these resources become widely used, as WordNet has become, they help inform and inspire hobbyists, high school students, and practitioners in industry who discover resources with strong, compelling theoretical foundations.

As they learn and understand the resource, these dilettantes become experts. They understand the quirks and practicalities of the resources often better than the original creators. Thus, it is important for the insights and the knowledge of users to be reflected in resources like WordNet so that resources can grow and thrive.

WordNet should be commended for being responsive to the community for so many years. However, now that technology enables such feedback to be directly, immediately, and efficiently integrated and redistributed into a linguistic resource, WordNet should, through its changing design, embrace and help the community to ensure that WordNet and resources like it continue to be useful, relevant, correct, and up-to-date.

References

[Boyd-Graber et al., 2007] Boyd-Graber, J., Blei, D. M., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of Emperical Methods in Natural Language Processing.*

[Boyd-Graber et al., 2006] Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted, connections to WordNet. In Sojka, P., Choi, K.-S., Fellbaum, C., and Vossen, P., editors, *Proc. Global WordNet Conference 2006*, pages 29–35, Brno, Czech Republic. Global WordNet Association, Masaryk University in Brno.

[Boyd-Graber and Resnik, 2010] Boyd-Graber, J. and Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Emperical Methods in Natural Language Processing*.

[Denecke, 2008] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. In *ICDEW 2008*.

[Hamp and Feldweg, 1997] Hamp, B. and Feldweg, H. (1997). GermaNet – a lexical–semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

[Kilgarriff, 2000] Kilgarriff, A. (2000). Review of WordNet : An electronic lexical database. *Language*, (76):706–708.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit. In *Tools and methodologies for teaching*.

[Miller, 1990] Miller, G. A. (1990). Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

[Ordan and Wintner, 2007] Ordan, N. and Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

[Russell et al., 2008] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77:157–173.

[Sagot and Fišer, 2008] Sagot, B. and Fišer, D. (2008). Building a Free French WordNet from Multilingual Resources. In *OntoLex*.

[Snow et al., 2006] Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808, Stroudsburg, PA, USA. Association for Computational Linguistics.