Reserating the awesometastic:

verb (transitive)

/ˌrɛzəˈreɪt/

To unlock; to open

adjective

/ɔːsəmtæstɪk/

superbly wonderful; fantastic, great

An automatic extension of the WordNet taxonomy for novel terms

David Jurgens and Mohammad Taher Pilehvar

jurgens@cs.mcgill.ca, pilehvar@di.uniroma1.it

network dynamics @mcgill

Which of these words is in WordNet?

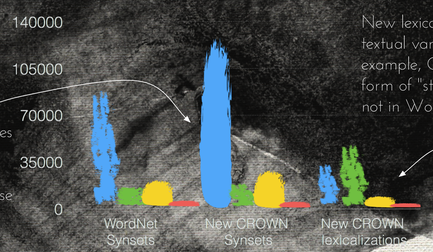
increment$_V$ remix$_N$ homie$_N$ homeboy$_N$ mp3$_N$ curate$_V$ benchmark$_V$ verifiability$_N$ microsoft$_N$ admin$_N$ fanbase$_N$ stepmom$_N$ unsubscribe$_V$ broadband$_N$ selfie$_N$ retweet$_N$ prequel$_N$ e-learning$_N$

Introduction and Motivation

As a semantic network and ontology, WordNet has been essential for many applications in Natural Language Processing. However, given the expense required to construct and update WordNet, it omits many domain-specific, slang, or recently-created terms — such as those shown to the left. While many works have proposed extending taxonomies such as WordNet, none have released any resources compatible with WordNet's existing tools or readily usable by the community. We present CROWN, an ongoing project that extends WordNet using an automatic enrichment procedure that integrates new terms from the community-curated content in Wiktionary. CROWN 1.0 more than doubles the number of new lemmas in WordNet and fills a crucial gap for researchers working in social media. Importantly, CROWN 1.0 is released in WordNet format, making it completely compatible with all existing WordNet tools and libraries.

Method and Results

Wiktionary glosses for terms not in WordNet are parsed and then matched against a series of heuristics to determine possible attachments. Heuristics include: a) recognizing people and taxonomic definitions, b) identify words antonymic to those in WordNet, c) using Wiktionary's limited thesaurus and structural information, d) identifying Wiktionary markup and e) identifying candidate hypernyms and finding the WordNet synset whose gloss has the highest semantic textual similarity.



Crown more than doubles the number of lemmas originally in WordNet without adding new sense distinctions for lemmas already in WordNet

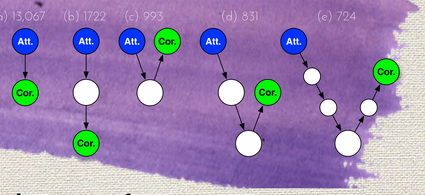
New lexicalizations allow recognizing textual variants as the same lemma. For example, CROWN includes "stepmom" as a form of "step-mom", the former of which is not in WordNet!

Experiment 1: Replicating WordNet

Methodology: To test the quality of data in Wiktionary and accuracy of our attachment algorithm, we replicate a portion of WordNet by re-attaching all monosemous lemmas in WordNet on the basis of their Wiktionary gloss. Error was measured in terms of how distant the estimated attachment was from the original taxonomic location.

Results: Attachments had a median error of three edges from the correct location, with most errors due to attaching to a more-specific synset.

Most common error patterns and frequency

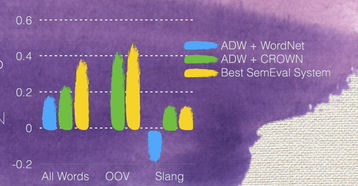


Example attachments

Experiment 2: Real-World Benefit

Methodology: To test the practical benefit of CROWN, we compare the performance of a state-of-the-art WordNet-based semantic similarity algorithm (ADW) when CROWN is used instead of WordNet. The algorithm was tested on a similarity benchmark based on SemEval-2014 Task 3, which includes slang and many terms not in WordNet.

Results: The use of CROWN improved the performance of the unsupervised ADW algorithm to nearly within that of current state of the art, which uses a significant number of resources. Thus, CROWN can be a highly-beneficial resource for WordNet-based approaches.



Fully Compatible With All WordNet Libraries

Check out the code and resource:

https://github.com/davidjurgens/crown

All of the code used to create CROWN is open-sourced and in on-going development!

Future work and SemEval-2016 Task!

CROWN is intended to be a continuously improved *community*-driven resource. Already, version 2.0 is underway with improvements such as new semantic relationships and integrating new senses of existing lemmas. We are organizing a SemEval-2016 task on taxonomy enrichment with the explicit goal of incorporating new insights directly into future versions of CROWN.