Proceedings of the Workshop on

Challenges for Wordnets

Francis Bond
Nanyang Technological University

bond@ieee.org

Maciej Piasecki
Wrocław University of
Science and Technology

maciej.piasecki@pwr.edu.pl

Language, Data and Knowledge Conference
Galway, Ireland, 18 June 2017

CEUR-WS.org/Vol-1899/wordnet_preface.pdf

Introduction: Comtemporary Challenges for Development and Application of Wordntes

Wordnets are increasingly widely used to model word meaning in natural language processing tasks. However, there are still many challenges in accurately describing word meanings and making these descriptions useful for both human and machine users. This workshop aims to identify, discuss and start to solve existing challenges. The idea grew from the CLARIN Workshop – Towards Interoperability of Lexico-Semantic Resources[1] held at Tartu (2017-01).

Wordnets are one of the most common semantic lexical resources and exist now for a large number of languages. Many are constructed manually or under careful manual control. Thus their quality is much higher than the quality of resources extracted automatically. Wordnets can provide a core for Linked Open Data that anchors them in language resources. Different wordnets are generally similar but still need some effort to combine in a common interoperable multilingual framework. It can be a first step and valuable case study in developing interoperable solutions for semantic lexical resources.

In addition to the Princeton WordNet of English, there are several large wordnets built for different natural languages, e.g. BulNet (Bulgarian), Czech Wordnet, DanNet, Dutch Open Source WordNet, enWordNet (English, a significant extension to Princeton WordNet), Estonian WordNet, Finish WordNet, GermaNet, MultiWordNet, plWordNet (Polish), sloWNet, as well as several semantic resources similar to or linked to wordnets. Many of them are very large and provide extensive coverage of vocabulary. All of them are somehow mapped to different versions of Princeton WordNet. There are also initiatives on building multilingual platforms linking several wordnets, e.g. Open Multilingual Wordnet. Some wordnets are connected to Linked Open Data that creates a possibility for anchoring named entities, their relational descriptions and ontologies in the natural language lexical semantics. Thus,

[1] https://www.clarin.eu/event/2017/clarin-workshop-towards-interoperability-lexico-semantic-resources

wordnets already form a rich set of lexical semantics resources that can be a case study in building interoperable solutions in semantic multilingual lexical resources for the needs of semantic interoperability on several levels: the infrastructure, meta-data, language tools and research applications. However this potential has not been fully explored yet. The main problems that we have identified before the workshop are:

- variety of formats for wordnets,
- small but significant differences between wordnet models that results in problems of mapping wordnet relations across language and makes their full use in multilingual applications more difficult,
- lack of a common set of applications for wordnet across different languages, e.g. for storing and accessing full content of wordnets (including their use as a reference set), text-to-sense mapping, word and text similarity calculation, etc.

The main challenges for wordnet developers seem to be:

- common formats for wordnets with their whole richness of semantic knowledge represented and models applied,
- inter-operability on the level of the interpretation of wordnet models (correspondence of basic elements and different types of relations),
- requirements for a basic set of wordnet-related tools, web services and applications,
- development of wordnet browsing services or inter-operable services,
- common framework for linking wordnets with Linked Open Data or domain knowledge resources for the needs of applications,
- gathering and sharing good practices in utilising wordnets and similar semantic resources in applications.

The program for this workshop was a little unusual. First, we asked participants to read the papers ahead of the workshop. Then the workshop itself was split into two sessions. In the first session, each participant had 15 minutes to highlight one or two of the challenges they considered most important, possibly with an outline of a solution. In the second session, we summarized the first session and then followed up with a long moderated discussion with all participating. Here we try to come up with a deeper understanding of the problems and sketch some possible solutions. Brief notes from the summary are included in this preface.

All the papers, the summary and notes from the discussions are online at: `http://clarin-pl.eu/workshop/wn-challenges.html`. We would like to thank CLARIN-PL[2] research infrastructure and the Global WordNet Association Board for their support.

[2] http://clarin-pl.eu

Organization

Organizers

Maciej Piasecki, *Wrocław University of Science and Technology*
Francis Bond, *Nanyang Technological University*
Jan Wieczorek, *Wrocław University of Science and Technology*

Program Committee

Eneko Agirre, *University of the Basque Country*
Francis Bond, *Nanyang Technological University*
Sonja Bosch, *University of South Africa, Pretoria*
Christiane D. Fellbaum, *Princeton University*
Darja Fišer, *University of Ljubljana*
Antoni Oliver Gonzalez, *Open University of Catalonia*
Shu-Kai Hsieh, *National Taiwan University*
John P. McCrae, *Insight Centre for Data Analytics, National University of Ireland,*
Galway Verginica Mititelu, *Romanian Academy*
Monica Monachini, *National Research Council of Italy*
Adam Pease, *Articulate Software*
Bolette Sandford Pedersen, *University of Copenhagen*
Maciej Piasecki (chair), *Wrocław University of Technology*
Alexandre Rademaker, *IBM Research*
German Rigau, *Polytechnic University of Catalonia*
Ewa Rudnicka, *Wrocław University of Technology*
Shikhar Kr. Sarma, *Gauhati University*
Stanisław Szpakowicz, *Emeritus Professor, University of Ottawa*
Veronika Vincze, *University of Szeged*
Piek Th. J. M. Vossen, *VU University Amsterdam*

Workshop Schedule

Time	Title	Presenter/Moderator
13:00–13:15	Welcome	Maciej Piasecki
13:15–13:30	How Stable are WordNet Synsets?	Eric Kafe
13:30–13:45	Inside Baseball: Coverage, quality, and culture in the Global WordNet	Martin Benjamin
13:45–14:00	Two corpus based experiments with the Portuguese and English Wordnets	Fabricio Chalub, Alexandre Rademaker and Cláudia Freitas
14:00–14:15	Overview and Future of Czech Wordnet	Adam Rambousek, Karel Pala and Sandra Tukacova
14:15–14:30	Challenges behind the data-driven Bulgarian WordNet (BulTreeBank Bulgarian Wordnet)	Petya Osenova and Kiril Simov
14:30–14:45	The Revision History of Estonian Wordnet	Neeme Kahusk and Kadri Vider
14:45–15:00	Towards Revised System of Verb Wordnet Relations for Polish	Agnieszka Dziob, Maciej Piasecki, Marek Maziarz, Justyna Wieczorek and Marta Dobrowolska-Pigoń
15:00–15:15	Classification of Adjectives in BulNet: Notes on an Effort	Tsvetana Dimitrova and Valentina Stefanova
15:15–15:30	The Concept of Lexical Platform	Maciej Piasecki, Tomasz Walkowiak, Ewa Rudnicka, Tomasz Naskręt and Francis Bond
15:30–16:00	Break	
16:00–18:40	Challenges for Wordnets: Summary and Discussion	Francis Bond (Moderator)
18:40–19:00	Conclusions, thank yous and farewells	Francis Bond and Maciej Piasecki

Discussion Notes

by Francis Bond (based on input from all present)

Outline These very rough notes are an attempt to list some of my (and Ed Hovy's) experiences; discuss a little what we are doing with the OMW; and summarize the main issues raised by the various papers. Views given here are not necessarily shared by all the participants (but we all seemed in general agreement). They were shown as slides to start the discussion and have been reproduced here with minimal changes.

- Some problems identified in Ontonotes
- Some problems identified by NTU-MC
- Some solutions with the OMW
- Summary
  - Technical
  - Social
  - Financial

Ontonotes Problems

- Technical
  - Missing senses
    named entities
  - Indistinguishable senses
    hard to distinguish reliably
    merged things until IA was 80%
    hard to tell if important information was lost
- Social
  - hard to add new terms

NTU Multilingual Corpus Problems

- Missing senses
  - especially MWEs and cultural words
  - non-english senses: *gohan* "cooked rice" *kome* "rice grains"

  - want to add multiple languages at once
  - small variations are common (sg/pl '-/space': *night bird* vs *nightbird*)

- Indistinguishable senses
  - distinguishing information may not be seen easily
  - verb frames, examples, network, …
  - some things (40 or so) are no distinguishable — merge

- Missing POS: pronouns, classifiers, greetings, exclamations, …

OMW Solutions

- Link projects with a common shared structure
  - document online and link to database (*)
  - link through the CILI
    also to other resources (SUMO, …)
  - allow multiple wordnets for the same language
    remove bottlenecks, encourage domain resources
  - more information about senses and variants
    allow orthographic variants, label, e.g. scientific names
  - link to corpora
  - rank using sense frequencies/automatically created ontologies
  - retain and publish statistics on usage and downloads (*)
  - allow online feedback and comments (*)
  - more tools for checking
    definition checking, suggest synonyms, panlex, …
    graph checks, corpus annotation, …

Permanent Identifiers

- Should we have sense-keys as well as synset identifiers?

- Will we use the lexicographer files any more?
  - Will PWN keep using these? (yes)

- How about version control for the whole wordnet?
  can we standardize on major.minor.patch?

Quality Control

- How can we get feedback for input regarding quality and selection?
    - user feedback
    - expert feedback
- Can we link to live examples?
- Can we copy structure more easily from other wordnets, …
- What is the best way to build a wordnet?
    - merge, transfer, some combination?
      automatic sources (from wiktionary, PanLex, …)
- Meanings change over time: how can we represent this?

Named Entities

- Should we have them or take them out?
- What about fictional names?
    - We need them in the Sherlock Holmes Corpus
- Can we get them from external resources
    - Geonames
    - Species
    - DBpedia
    - UMLS

Domains and Productive Meaning Extensions

- Domains
    - Should we move them into sub wordnets?
    - Or just add better domain markers?
    - Should we keep supertypes (lexicographer files)?
- Productive Meaning Extensions
  how should ee handle these
    - *quick* to *quickly*
    - *actor* to *actress*
    - *interest* to *interested*
    - *clasp* to *unclasp*

Definitions and Examples

- It is hard to write a good definition? Yes
- Should it be unique? Yes
- Should it have redundant info from the links (e.g. domain)? Yes
  possibly show in different interfaces
- Should it be disambiguated? Yes
- What do we do if it changes?

Ontology

- Can we use wordnet as an ontology? No
- What are the differences between WN and Ontologies?

Compatibility

- The same thing is called something different by different groups
  - project to a shared vocabulary (OMW)
  - document this (calling for volunteers)
- Several solutions to similar problems
  - lexical gaps, phrasesets, artificial synsets,
- Mapping
  - should we allow a variety of links to CILI?
  - not just equal-synonymy

Social

- Restrictive licenses are restrictive
  - Unless you can redistribute the data, it is hard to make a large compatible collection, …
  - Unless the licenses are compatible it is illegal to merge, …
- How do we get feedback from end users/applications?
  e.g. If Kamusi adds something, how does it get back to the WN

- How do we get acknowledgments back to the original wordnets
  e.g. If someone looks it up in OMW/NLTK/Babelnet how is credit fed back?

- How do we know what is happening with wordnets/roadmap?
  Should we discuss this at GWC?

- We don't have standard tools that we all use
  so duplication of effort
  can we have a fixed data-specification so we can use any tool

Financial

- It is hard to get money to maintain wordnets
  - How can we make things easier to maintain?
- How can we persuade funders?
- How can we use funding/effort most effectively?