Extending Wordnet to Geological Times

Henrique Muniz (FGV/Emap)

Fabricio Chalub (IBM Research)

Alexandre Rademaker (IBM Research and FGV/EMAp)

Valeria de Paiva (Nuance Comm.)

Wordnet Conference 2018 - Singapore

Motivation

- Users should be able to query a document database for retrieving documents that mention some concepts of interest.
- Some queries embody implicit knowledge.
- Using a domain ontology we can enrich the query and retrieve more documents of interest. Query expansion problem.
- Examples:
  - If the query mentions “Intraplate basin”, it is possible to retrieve also documents that do not mention explicitly “Intraplate basin”, but that mention “intracratonic basin”, because “intracratonic basin” is a subtype of “intraplate basin”.
  - The same can be done with lithologies. A query mentioning ”sedimentary rock” can retrieve documents that mention ”sandstone”, because ”sandstone” is a subtype of “sedimentary rock”.
- Requirements:
  - The ontology should be able to capture the relevant concepts that will be used in the queries. Lithologies, image patterns, notions related to sedimentary basin and petroleum systems.

Searching for “intraplate” should retrieve “intracratonic” basin?

1. Particularly revealing is the window opened by the presence of abundant molartooth structure onto the paleoceanography, paleobathymetry, paleoclimate and tectonic regime of this **intracratonic Precambrian basin**.
2. The Williston Basin is an **intracratonic sedimentary basin**, with the Nesson and Cedar Creek anticlines as major structures in the basin (Figure 1).
3. The Reggane basin is an **intracratonic Palaeozoic basin** located in the south-western Saharan platform with an area of 140 000 km².
4. The Paleozoic Hudson Bay Basin in northern Canada is one of the largest, yet least explored **intracratonic basins** in North America.
5. The breakout orientation for **both intracratonic and pericratonic areas** of KGB varies from N36"W to E-W corresponding to geologic ages of sediments from the Permo-Triassic through Miocene.

Steps

1. The English Slot Grammar Parser (IBM grammar-based parser)
   - intraplate basin = intra+plate as adj modifying basin
   - intracratonic basin = intra+cratonic as adj modifying basin

2. Wordnet Lookup (no relation between plate and craton)
   - plate = http://wnpt.brlcloud.com/wn/synset?id=09395457-n (hyponym of crust)
   - craton = http://wnpt.brlcloud.com/wn/synset?id=09259500-n (hyponym of piece, part of continent, is an old and stable part of the continental lithosphere)
   - cratonic *derivated related to* craton
   - basin = http://wnpt.brlcloud.com/wn/synset?id=09215437-n
   - intra (?)
3. What do we need to infer?
   - intracratonic Precambrian basin < intracratonic basin
   - intracratonic sedimentary basin < intracratonic basin
   - intracratonic Palaeozoic basin < intracratonic basin
   - Basin is an area
   - and “intracratonic basin” is a subtype of “intraplate basin”

English Slot Grammar

```
It reviews the initial production methods utilized to control
water and sand production.
------------------------------------------------------------------------
.------------ subj(n)      it(1)                   noun pron sg def perspron
o------------ top          review1(2,1,6)          verb vfin vpres sg vsubj ansb vthink (nform review) (ernform reviewer)
| .----------- ndet         the1(3)                 det pl def the ingdet
| .----------- nadj         initial1(4,6)           adj
| .----------- nnoun        production1(5,u,u,u,u)  noun cn sg evnt act abst transaction groupact comm (latrwd 0.061050) (vform produce)
`-+----------- obj(n)       method1(6,u,u,u)        noun cn pl cognsa (latrwd 0.115380)
  `----------- nnfvp        utilize1(7,u,6)         verb ven vpass vlast (nform utilization utility) (ernform utilizer)
    `--------- vnfvp        to1(8,9)                infto
      `------- tocomp(binf) control2(9,u,13,u)      verb vinf (nform control) (ernform controller)
        |   .- lconj        water1(10,u)            noun cn sg locn massn sbst ent wlocn (latrwd 0.113050)
        | .-+- nnoun        and1(11)                noun cn sg pl sgpl massn sbst cord ent
        | | `- rconj        sand1(12,u)             noun cn sg massn sbst material ent (latrwd 0.032630)
        `----- obj(n)       production1(13,u,u,u,u) noun cn sg evnt act abst transaction groupact comm (latrwd 0.061050) (vform produce)
------------------------------------------------------------------------
```

Surfasce and deep structures; Multiwords and named entities; Lexical augmentation via Wordnet; Logical forms; PTB output (dependencies to constituents); subject areas and features; many languages via LMT translations; (first paper back to 1980)

M. C. McCord, J. W. Murdock, and B. K. Boguraev, “Deep parsing in Watson,” IBM J. Res. & Dev., vol. 56, no. 3, pp. 3:1–3:15, May 2012.

English Slot Grammar

```
Precambrian carbonate rocks are abundant in the intracratonic Sao
Francisco Basin composed mostly by microbial facies.
------------------------------------------------------------------------
  .----------- nadj        Precambrian1(1,3)        adj capped
  .----------- nnoun       carbonate1(2,u)          noun cn sg massn sbst ent (latrwd 0.023080)
.------------- subj(n)     rock1(3,u)               noun cn pl physobj massn sbst natent ent material (latrwd 0.159870)
o------------- top         be(4,3,5)                verb vfin vpres pl vsubj absubj auxv
`------------- pred(a)     abundant1(5,3,6)         adj
  `----------- aobj(p)     in1(6,5,8)               prep staticp
    | .------- ndet        the1(7)                  det sg def the ingdet
    `-+------- objprep(n)  intra+cratonic1(8,u)     noun cn sg
      `------- nprop       Sao Francisco1 Basin(11) noun propn sg glom capped
      `------- nnfvp       compose2(12,14,8,u)      verb ven vpass (nform composure composition composing) (ernform compositor composer)
        `----- vadv        mostly1(13,12)           adv
        `----- subj(agent) by1(14,12,16)           prep pprefv
          | .- nadj        microbial1(15,16)        adj
          `--- objprep(n)  facies1(16,u)            noun cn sg pl sgpl
------------------------------------------------------------------------
```

```
    Precambrian(e1,x3) pos(e1,JJ)
    carbonate(e2,x2,u) nmod(e17,x2,x3) pos(e2,NN)
  rock(e3,x3,u) pl(e3) pos(e3,NNS)
be_adj(e4,x4,x3,e5) vpres(x4) pos(e4,VBP)
  abundant(e5,x3,e6) pos(e5,JJ)
    in(e6,e5,x8) pos(e6,IN)
       the(e7,e8) pos(e7,DT)
      intra+cratonic(e8,x8,u) pos(e8,NN)
        Sao\ Francisco\ Basin(e11,x11) pos(e11,NNP)
        compose(e12,x12,x16,x8,u) pos(e12,VBN)
          mostly(e13,e12) pos(e13,RB)
            microbial(e15,x16) pos(e15,JJ)
          facies(e16,x16,u) pos(e16,NNS)
```

Examples of sentences and title

- Rock mechanics tests on core from Early Cretaceous carbonate reservoirs from a super-giant field offshore Abu Dhabi has allowed definition of rock mechanical facies (RMF).
- Oceanography, bathymetry and syndepositional tectonics of a Precambrian intracratonic basin: integrating sediments, storms, earthquakes and tsunamis in the Belt Supergroup (Helena Formation, ca. 1.45 Ga), western North America
- The Big Lime in the Dickenson field is a stratigraphic trap where porous oolites laterally grade into nonporous, nonoolitic limestone.
- Natural gas production in central West Virginia is primarily from the shales of Devonian age upward to the sandstones of Pennsylvanian age. (ambiguity in the parsing)

Mineral vs rock type or both?

- The rock: https://en.wikipedia.org/wiki/Dolostone

  The term *dolostone* was introduced to avoid confusion with the mineral *dolomite*. The usage of the term *dolostone* is controversial because the name dolomite was first applied to the rock during the late 18th century and thus has technical precedence. The use of the term dolostone is not recommended by the *Glossary of Geology* published by the American Geological Institute. It is, however, used in some geological publications.

- The mineral: https://en.wikipedia.org/wiki/Dolomite

  The term is also used for a sedimentary carbonate rock composed mostly of the mineral dolomite. An alternative name sometimes used for the dolomitic rock type is dolostone.

- In PWN dolomite it is a hyponomy of Rock (http://wnpt.brlcloud.com/wn/synset?id=14838055-n)

GeoScience Natural Language Processing Pipeline: the TKB

- Hope to evolve to initiatives such as “Open Health Natural Language Processing Consortium” http://www.ohnlp.org/index.php/Main_Page
- The English Slot Grammar
- Prolog for postprocessing the parse trees. Alternative to Jape/Gate and AQL (Sytem T, IBM Almaden)
  - L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!,” Conference on Empirical Methods in Natural Language Processing, Seatle, Washington, 2013, pp. 827–832.
- WSD using UKB (shallow/deep processing!?)
- Ontologies: GeoNames, ISC etc); OpenWordnet-EN (PWN+) etc.

Are extensions of PWN really necessary?

- Issues related to tokenization: words like *ping-pong, kickboxing, water-ski* and *fistfight* should appear with space, hyphens or not, in the respective synsets. Quite a bit of post-processing is necessary.
- It would be good to add many prefixes, suffixes and regular endings, which are perfectly understandable by humans, but not so much by machines: s*hirtless* and *localizer, focalizer* are not in WordNet. Also many verbs ending in *-ize, -ise* or *-ify* are not present in PWN, while being in Wiktionary, for instance *adjectivise, Africanize* or *incentify, girlify*.
- Many other problems already reported in previous presentations!

Are extensions of PWN really necessary?

- We want to use WordNet as a sort of “lightweight” ontology
- Some shallow reasoning can be done on the basis of lexical taxonomies => best to cover all concepts, at the expense of being shallow than to have big gaping holes in the concepts covered.
- boot-strapping specific domain ontologies for specific domains.
- geological concepts, need to be fitted within the taxonomic framework of a lexical knowledge base like WordNet, so that we can take advantage of the aforementioned framework.
- Starting with a subset of Geological domain: WordNet specific information concerning **geological time periods**.

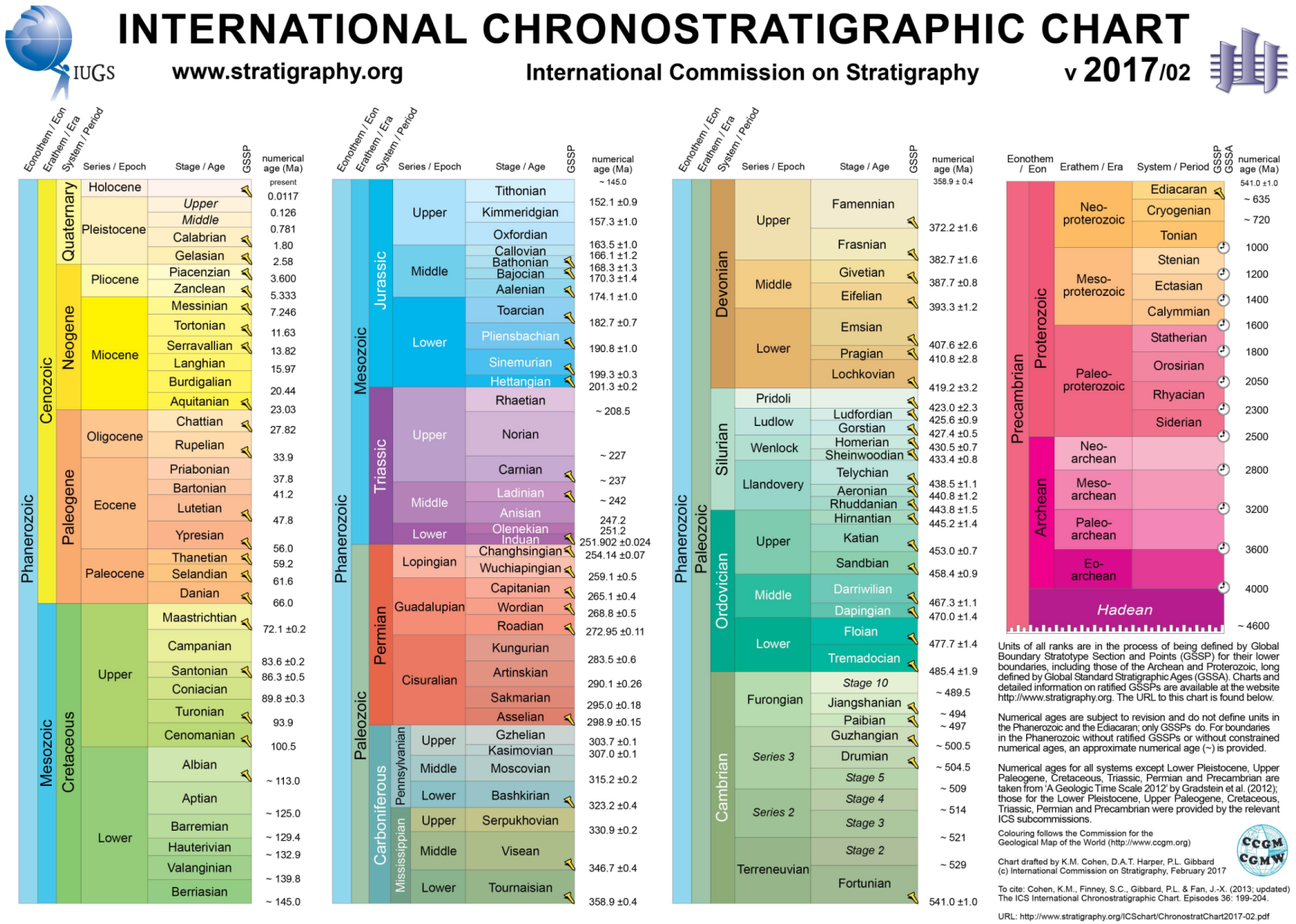
WN for GeoScience

- WordNet is concerned not inflating the lexicon with terms that are clearly derived
- (e.g.*coaly* is simply the adjective form of having to deal with *coal*) or
- easily compositional (like *basinward* – in the direction of a basin).
- new expressions consisting of prefixes and suffixes are not considered, WordNet has *aeon*, but not *super-aeon*.

WN for GeoScience

- While the common noun *stratigraphy* is in PWN, even the adjective *stratigraphic* is not in the database and neither is the compound *chronostratigraphic*.
- Prefix *chronos*, meaning ‘time’ and the suffix denoting a pertainym adjective *–ic*
- *Strata (plural of stratum) is in the gloss but not in the lexical unit.*
- We would like to devise and describe a process to extend WordNet for a specific domain
- We use geological time periods and a small collection of papers in Petrology as a paradigmatic example of a domain specific extension.

International Commission on Stratigraphy



GTP are not as well-established

Most of the systems, series and stages were first defined from type-sections in Europe, the historical home of stratigraphy.

Subsequent study of stratigraphical successions worldwide has led to a proliferation of regional units. These historical units did allow Phanerozoic strata to be correlated and mapped world wide.

However, as it happened, most successive chronostratigraphic units are located in geographically separated type sections, which have more recently been shown to be separated by significant gaps or to overlap considerably.

These problems, and the general lack of defined boundaries for historically established units, became serious hindrances to high-resolution correlation of geographically widespread stratigraphic successions.

GTP in PWN

- The chart mentioned above contains 176 names of geological periods. Of these only 28 are in WordNet and all but 40 are in Wiktionary. The last 11 are in Wikipedia, but not in WordNet or Wiktionary.
- Hyponyms of [15116283-n:*geological time, geologic time-the time of the physical formation and development of the earth (especially prior to human history)*].
- Hyponyms include synsets for each of *aeon, geological era, geological period* and *epoch*.
- Eons (or aeons) are divided into eras. Eras contain periods that contain epochs, and finally epochs contain ages.
- The first three eons (Hadean, Archean, Proterozoic) are collectively referred as the Precambrian super-eon.
- The most recent eon, the Phanerozoic is subdivided into several periods. All of these five names of periods have their respective synsets, but *super-eon* is not in PWN.

How to extend PWN?

- One reasonable way for a specific field is to process a corpus of quality texts in this field and check for missing entries.
- Another avenue of expansion was to incorporate a domain-specific ontology created by the professionals of the area. The ISC Ontology.
- Not enough, discovering when compounds are to be treated as multiword expressions, as opposed to compositional compounds, is a challenge.

  GSSP (Global Boundary Stratotype Section and Point) is called "*golden spike*" *(*an internationally agreed upon reference point on a stratigraphic section which defines the lower boundary of a stage on the geologic time scale)

The ISC ontology

- *age, eon, epoch, era, period, sub-period*, and *super-eon* are sub-classes of GeochronologicEra (abbreviated as GE).
- However, there is no formally defined hierarchy between these concepts.
- Greater emphasis is placed on the boundaries of the periods and only the approximate duration of the period is given in the chart.
- It is important to note that geologists qualify the units as “early”, “mid”, and “late” when referring to time, and “lower”, “middle”, and “upper” when referring to the corresponding rocks.

  lower Jurassic Series in chronostratigraphy corresponds to the early Jurassic Epoch in geochronology

The ISC Ontology

- The boundaries between periods are annotated using another ontology, the Temporal Hierarchical Ordinal Reference System model(THORS4)
- The time interval of a GE is given in terms of its boundaries to other GEs via **thors:begin** and **thors:end**.
- Each boundary is a **GeochronologicBoundary** and it is temporally located via **iso19108:temporalPosition** which specifies a **iso19108:Coordinate** with a value, frame (e.g., “Ma”), and a positional uncertainty.

Example

- Maastrichtian period in Wikipedia : “in the ICS geologic timescale, the latest age or upper stage of the Late Cretaceous epoch or Upper Cretaceous series, the Cretaceous period or system, and of the Mesozoic era or erathem”.
- 176 basic geologic period terms is easy to deal with.
- However, we still need common nouns (*play, basin, cleats*) and compounds (*golden spike*), whose geological meanings are very different from their usual meanings.
- These need to be extracted from a geology corpus.

```
Maastrichtian a GeochronologicEra ;
  rank Age ;
  begin BaseMaastrichtian ;
  end BaseCenozoic .
BaseMaastrichtian a GeochronologicBoundary ;
  temporalPosition BaseMaastrichtianTime .
BaseCenozoic a GeochronologicBoundary ;
  temporalPosition BaseCenozoicTime .
BaseMaastrichtianTime a Coordinate ;
  frame ma ;
  value "72.1" .
BaseCenozoicTime a Coordinate ;
  frame ma ;
  value "66" .
```

Corpus

- 155 randomly selected passages relevant to petroleum systems extracted from 1,298 publicly available English language geological reports.
- The passages were segmented in 5,661 sentences (186,244 tokens) and parsed in the Universal Dependencies scheme by Udpipe.
- Out of 8800 noun lemmas uncovered by UDpipe, more than half were not recognized as present in WordNet.
- the corpus is full of named entities, that cause Named Entity Recognition to be a hard task.
- Some of these missing words are processing mistakes. ‘reservoirs’ was not correctly lematized to ‘reservoir’.
- A large proportion are named entities that WordNet is not supposed to list.
- But a small proportion are really common words that WordNet should have. Finding these seems to be a positive side e-fect of trying to extend WordNet for specific domains.
- we look at all with more than 10 occurrences

Creating Synsets in PWN+

- As we want to map all their precision into an extended version of Princeton WordNet we need a kind of a *domain specific language*(DSL) to describe new synsets.
- Improved lexicographer files. Version control; easy edition and localization of descriptions; mainly for human consumption; Redundancies are eliminated; Artificial ids are avoided to make maintenance easy; mnemonics (hyper, ant) instead of symbols for relations.
- To maintain compatibility with existing systems that already use PWN sense keys and synset ids we provide mappings between our sense ids and PWN. Similarly, mappings that link synsets and existing ontologies can also be defined.
- https://github.com/own-pt/own-en/blob/master/dict/noun.geotime.txt

Using PWN+

In this chapter, the kinematic interpretation of the west Carbonate shear zone is placed in a regional context, with regard to intrusive and tectonic activity from 2740 to 2690 Ma ago.

Assuming the parser got it right and the range and units are detected.

Query returns three ISC entries: the Neoarchean era (2500–2800Ma), the Archean eon (2500–4000 Ma), and also the Precambrian super-eon (541–4567 Ma).

```
select ?era ?rank ?vbegin ?vend
{
 ?era gts:rank ?rank ;
   thors:begin ?tb;
   thors:end ?te .

  ?tb ts:temporalPosition ?begin;
  ?te ts:temporalPosition ?end .

  ?begin ts:frame age:ma ;
     ts:value ?vbegin .

  ?end ts:frame age:ma ;
     ts:value ?vend .

  bind (2690 as ?a)
  bind (2740 as ?b)

  filter ((?a <= ?vbegin &&
           ?a >= ?vend) ||
          (?b <= ?vbegin &&
           ?b >= ?vend))
}
```

Future work

- Integrate with TKB improve the pipeline.
- Alternatives do UKB for WSD or integrated with the grammar.
- Add more extensions
- Working with Glossaries for the domain
- Further experiments
- Mapping for other ontologies and contribution to SUMO for formal definitions and evaluate the benefits.