MSDS 7337 - NLP - Glossary

Patrick McDevitt[1]

Masters of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
pmcdevitt@smu.edu

Abstract. This glossary is a catalog of interesting terms encountered in the study of the SMU NLP course during fall semester 2018.

1 Unit 1

- NLP - natural language processing - NLU + NLG
- NLU - natural language understanding - We try to get the machine to produce a useful representation of some inputted natural language
- NLG - natural language generation - We try to get the machine to produce usable, natural language output that is not just identical to its input.
- Applications of NLU
  - Text annotation - tagging, metadata extraction/generation, classification, document summarization
  - Corpus analytics - theme extraction, clustering, taxonomy mapping, sentiment analysis
  - Search applications - query repair, query refinement, results postprocessing (ranking, clustering, encapsulation)
  - Advanced applications - machine translation, knowledge discovery, question handling
- Applications of NLG
  - Text annotation - document summarization, generation of callouts / headlines
  - Corpus analytics - labelling of clusters, synopsizing of corpus-wide topic and/or sentiment trends
  - Search applications - advanced capsule generation (summarization modified to fity the query), advanced query refinement (next gen version for disambiguation)

- Advanced applications - machine translation, knowledge discovery, question handling (refinement, answering)

2 Unit 2 - Levels of Analysis in NLP

– Levels of NLP analysis
  - Lexical Analysis – Words
    * lexicon - dictionary, vocabulary
    * morphology - study of morphemes - units of which words are made
    * stemmer - algorithm that provides root morpheme
    * metadata - corpus derived metadata supports developing : word frequency score, common collocation, and commonly co-occurring words and context words
    * collocations - words commonly occuring together
    * head word - specific listing in the lexicon
    * polysemous - more than one meaning
    * WordNet - the Bible of word senses, produced by Cognitive Science Laboratory of Princeton University - Psychology Dept -, shows relationships between words. many real world senses are not in WordNet (i.e, apocryphal)
    * domain relations - mode of discourse around certain activity or subject matter
    * Uses of lexical analysis - text and corpus analytics : spell correction, terminology extraction (OCR), lexical diversity measurement
  - Syntactic analysis – Grammar
    * sentence boundary detection - needed since syntax analysis is sentence by sentence analysis. grammar parsing relies on sentence boundary detection. non-trivial problem.
    * part of speed (POS) - each word in a sentence can be assigned to a part of speech. most words have more than one part of speech; therefore, POS tagging is non-trivial. needs contextual clues - words preceding and succeeding,
    * Penn Treebank Tagset - commonly used tag sets for parts of speech, approx 40 different POS with this set, support eventual grammar parsing
    * parsing - break the sentence into grammar parts

    * Lemmatizaton - canonical (conventional) form that represents a set of related word forms
    * discrete text field analysis - unitizing, normalizing, smart ETL.
- Semantic analysis – Meanings
    * named-entity extraction - NEE or NER - recognize entity without typing - persons, organizations, places, events. good NEE will cluster together many variants, including epithets. not as simple as looking up entities in wikipedia
    * relationship extraction - need syntax analysis and semantics - need a representation of the world (an ontology)
    * - word sense disambiguation - still unsolved problem in AI - how to distinguish the meaning of a word in context, from all of the possible meanings of the word
    * classification - use a tree-structured graph to place documents into categories - often use machine learning (such as SVM)
    * taxonomy - hierarchical structure in which each entity is assigned to one category. IAB is likely most influential taxonomy - Internet Advertising Bureau - is huge part of internet experience.
- Discourse / Entailment Analysis – Inferences