MWE and NE handling in DELPH-IN

Alexandre Rademaker

July 16, 2020

Contents

1 introduction

1.1 definitions

A multiword expression (MWE) is a combination of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies. MWEs encompass diverse linguistic objects such as idioms (to pull the strings 'to make use of one's influence to gain an advantage'), compounds (a hot dog), light-verb constructions (to pay a visit), rhetorical figures (as busy as a bee), institutionalized phrases (traffic light) and multiword named entities (European Central Bank). ref

- NE are MWE
- proper names vs named entities
- titles of works (films, books) are named entities and MWE (?)

1.2 motivation

- grammar coverage
- information extraction, the semantic representations (e.g. ‘padding pool‘)

1.3 approaches

- What is the DELPH-IN approaches for MWE and NE?
- pre-processing (PET input formats)
- post-processing (wordnet)
- combination of deep vs shallow. See `http://heartofgold.dfki.de` (PET inputs!)
- the influence on the parsing cost
- semantic representation simplicity and consistency

2 previous discussions

- `http://moin.delph-in.net/TomarNames` (2014)

  How to parse named entities constructions. Inconsistencies in annotations.

  Dan's comment "...capital letter specifically flags opaque proper names. But that leads to a radically different semantics in our representation." about unknown/generic with _[lemma]$_{\text{nrel}}$ vs CARG

  I'm interested in exploring is having the predicate name in all nouns be _[lemma]$_{\text{nrel}}$ for State or state, while recording the capitalization somewhere in the EP, probably on the ARG0.

  "The decision about conflating the representations of named and non-named things will address a large class of difficulties I encountered."

  Proposed grammar behavior: In common noun case, only a single lexical entry with an annotation from the preprocessor, then a non-branching rule sensitive to the property. Trying to avoid having two lexical entries Justice and justice in the chart, so annotators don't have to decide.

Dan "If we build more sophisticated pipelines with either good preprocessors or post-processing of the MRSs. What I'm aiming here to do is to take out our pretense of being able to make that decision sentence-internally with high reliability."

Emily: I think you've been using proper name in this discussion where I'd use named entity.

- `http://moin.delph-in.net/ErgSemantics/Essence` (2018)

  We view a comprehensive and consistent treatment of named entities to be an open and challenging research topic, meriting a detailed discussion outside the scope of this overview.

  BTW: ‘President Smith‘ why ‘$_{\text{presidentnof}}$‘? What ‘of‘ means?

- `http://moin.delph-in.net/PetInput` (2012)

  the old code has been augmented over time with additional procedural mechanisms, all aiming to 'transport' token-level surface properties into the grammar-internal feature structure universe. Examples of such mechanisms are so-called characterization (recording of string-level start and end positions for each token) and the determination of CARG and PRED values in the MRS component of grammar-internal feature structures, in both cases reflecting the token surface form of named entities or predicates introduced by other generic entries.

  . . . a set of token mapping rules prior to lexical instantiation looks for string-level indicators of various kinds of NEs

- `http://moin.delph-in.net/PetInputChart` (2011)

  It extends the YY mode in that it allows to have structured input tokens to provide a means to encode, say, named entities resulting from base tokens.

- `http://moin.delph-in.net/SmafTop` (2011)

  Properties of each edge: . . . a type (eg. token, pos, namedEntity, morphosyntax, . . .)

- `http://moin.delph-in.net/ErgProcessing`

  lightweight RE-based named entity detection

- `http://moin.delph-in.net/CambridgeEfficiencyRobustnessPrecision`

Francis: If one of the things we do is produce more training data, one approach is to process other people's markup. e.g. in WikiWoods, use wikilinks to produce Dan's special **constituent brackets**.

- `https://delphinqa.ling.washington.edu/t/using-erg-for-information-extraction/177/2`

  Simple proper nouns, like "Fassett" in your example, are named EPs quantified by $\text{proper}_{\text{qs}}$. Slightly more complicated are compounds, like "La Ventana", where a compound EP joins two named EPs (and in this case, the syntactic head noun is quantified by $\_\text{the}_{\text{q}}$ instead of proper_{q}). Coordinated names ("Bill and Melinda Gates") are more complicated. Some proper nouns include common nouns, such as "The University of Washington", which has $\_\text{university}_{\text{n1}}$ and not named("University"), so the line starts to blur about where the proper nouns begin and end.

- `http://moin.delph-in.net/MweTop` (2013)

  - `http://www.lrec-conf.org/proceedings/lrec2002/pdf/145.pdf`

    A guiding principle is that, where possible, MWEs should be related to simplex entries. Ongoing work involves refining the formal representation of the MWE classes and deciding on database structures.

  - `https://dr.ntu.edu.sg/bitstream/10220/6828/1/2002-cicling-mwe.pdf` (Multiword Expressions: A Pain in the Neck for NLP) (2002)

    "Proper Names are syntactically highly idiosyncratic." Classified as a kind of Semi-Fixed Expressions.

    Therefore, the constraint on Name is defeasible: it can be overridden in rules that inherit from it. The logic for defaults we assume follows Lascarides and Copestake (1999).

    The sentence 'The Oakland Raiders won the game.' was analysed as suggested in this paper? Is it implemented?

- `http://moin.delph-in.net/SaarlandMweDiscussion`

  "Discussion of MWEs, inspired by Ann's participation in PARSEME." Not much structure.

- `http://moin.delph-in.net/HeGram`

  Some works on verbal MWE constructions.

- http://moin.delph-in.net/SingaporeRepresentingMwes

  Slides not available: http://lingo.stanford.edu/delphin2015/isf-poss.pdf

  Emily: $\mathrm{id}_{\mathrm{rel}}$ replaced with ICONS?

  Francis: Have played some with pyDelphin; also talked with Ann about her packed DMRS representation.

  Tuan Anh: Imagine that you are a human translator who had to translate this sentence.

  conjunctive packed MRSs?

- http://moin.delph-in.net/DelphinLingo

  The ERG has been used in research on multiword expressions (MWEs) jointly funded by the NSF and NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, and the ERG was also extended for use in hybrid information extraction as part of the Deep Thought project.

  http://heartofgold.dfki.de ?

- https://www.aclweb.org/anthology/W15-0128.pdf (Layers of interpretation. . .) (2015)

  Turning to types of semantic annotation which are not compositional, we first find layers that concern only atoms. These include fine-grained word-sense tagging, named entity tags and so on. According to the definition we have given, there may be an indefinite number of atom-meaning pairings, but these are outside the scope of the compositionality principle.

- http://moin.delph-in.net/ErgSemantics/Inventory

  MWE Quantifiers: ‘any more‘ et al – not documented

- http://www.coli.uni-saarland.de/~yzhang/files/delphin2007-zhang.pdf

  Lexical Acquisition for MWE Heads (compositional approach) vs "words-with-spaces" improves coverage.

  Hypothesis: the relative ordering in frequency for different n-grams is preserved across corpora, in the same domain. If not, different conclusions may be drawn from different corpora

3 Universal Dependencies

- MWE can be ‘fixed‘, ‘flat‘ or ‘compound‘

  fixed fixed grammaticized expressions that behave like function words or short adverbials. They are the fixed expressions category of Sag et al.

  flat exocentric (headless) semi-fixed MWEs like names (Hillary Rodham Clinton) and dates (24 December). But names that have a regular syntactic structure, like "The Lord of the Rings", should be annotated with regular syntactic relations. Organization names with clear syntactic modification structure, the dependencies should also reflect the syntactic modification structure.

  compound applies to endocentric (headed) MWEs (like apple pie): particle verbs, serial verbs, noun compounds.-

- `https://universaldependencies.org/u/dep/flat.html`
- `https://universaldependencies.org/u/dep/compound.html`

some open issues about named entities in docs. For example, using ‘features’ for tag names, confusion about relations or format.

4 PARSEME

- discussions about verb expressions, clitics etc
- data is a different layer of annotation on top of UD
- for LVC differences between parseme and propbank.

5 AMR

- Limitation in guide line - it does not deeply capture many noun-noun or noun-adjective relations.
- many extensions
- any concept can have name relation - named entities
- entity links with ‘:wiki‘
- deep structure with ‘:mod‘, ‘:location‘ etc