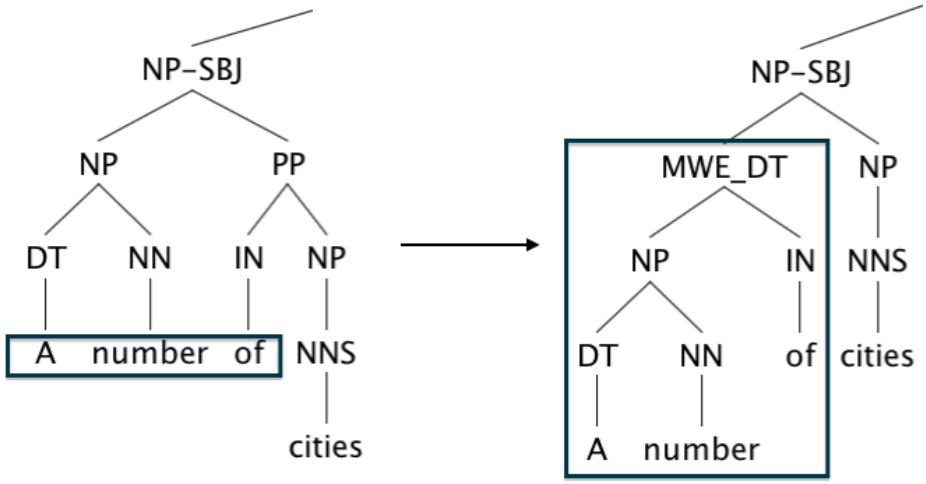
MWE-aware English Dependency Corpus (Version 2.0)

We provide users with an English dependency corpus taking into account multiword expressions (MWEs) built on the Wall Street Journal portion of Ontonotes Release 5.0 (LDC2013T19). Among various kinds of MWEs, we deal with named entities (NEs) and compound function words, which serve as functional expressions.

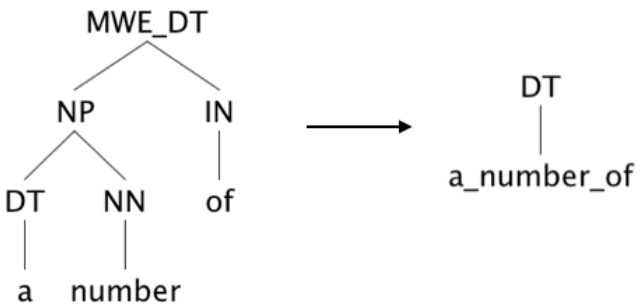
Corpus Construction

We built the corpus according to the following steps [1][2].

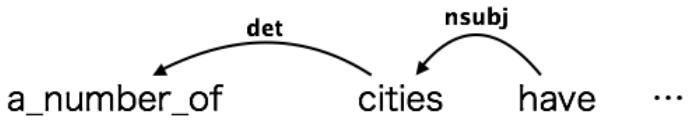
1. We find an MWE in the phrase structure trees of Ontonotes and establish it as a single subtree.
    - Regarding compound function words, we utilize MWE-spans and MWE-level POS tags provided by [3].
    - Regarding NEs, we utilize NE annotations provided by Ontonotes release-5.0.
    - The phrase structure trees made by this step are also provided.



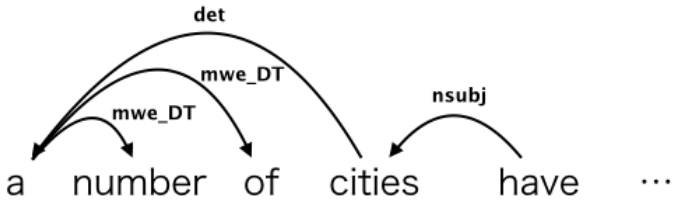
2. We replace the above subtree by a preterminal with its leaf node as a child. The preterminal has the same part of speech as that of the MWE. Its child node is made by joining all components of the MWE with underscores.



3. We convert the phrase structure into Stanford Dependency Ver.3.5 [4].
    - We designate "-conllx -basic -makeCopulaHead -keepPunct" as a option for the conversion command.
    - An example of an MWE-aware Dependency tree is given in the figure below.



4. We decompose the token derived from an MWE (e.g. a_number_of) to a ``head-initial'' dependency structure taking into account the consistency with Universal Dependency [5]. In other words, each token of an MWE modifies the first token. We use special dependency labels which start with "mwe" followed by an MWE-level POS tag (e.g. mwe_RB, mwe_IN).



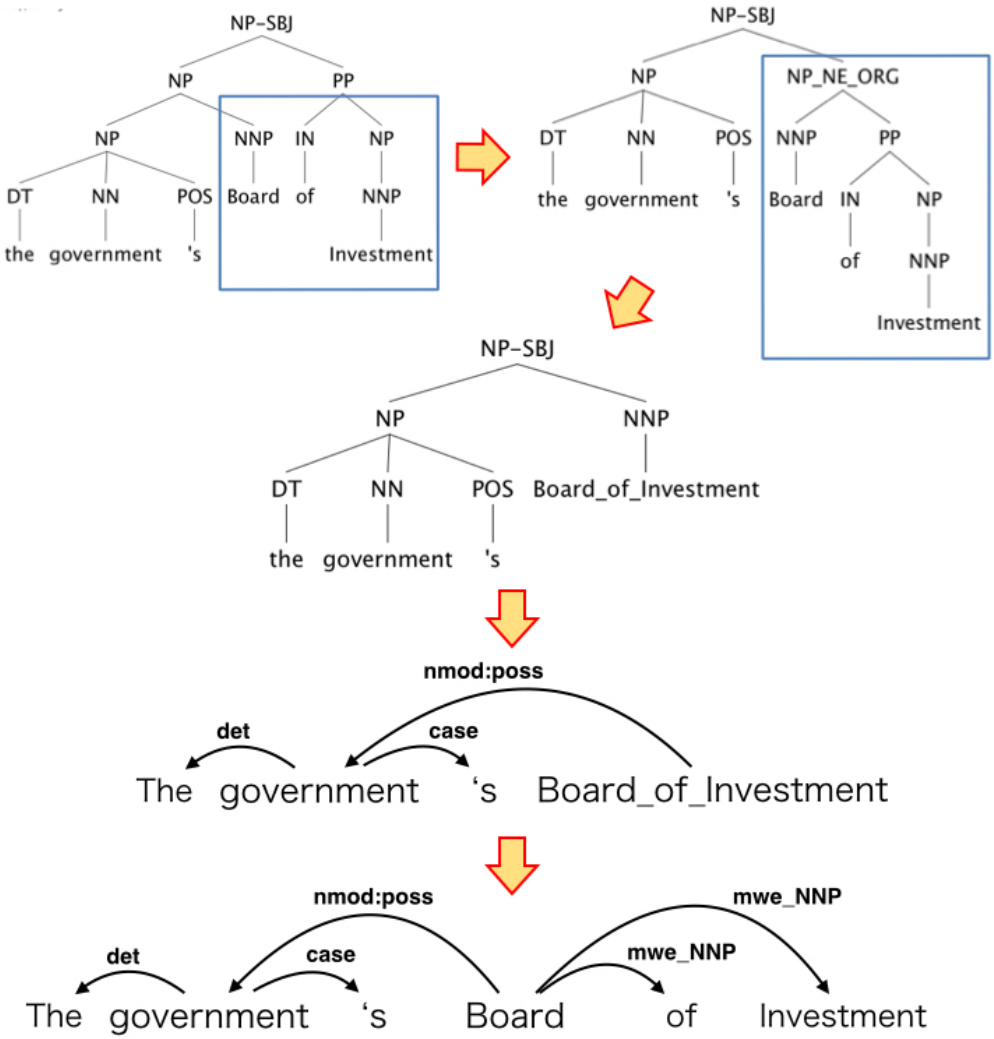
Additional Note regarding Named Entities

We exploit NE annotations on Ontonotes Release 5.0 (LDC2013T19).

We address traditional NEs, such as persons, locations and organizations, while not dealing with the followings: DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL. Note that we focus on only multiword NEs.

For some instances, it is more reasonable to enlarge NE-spans than to modify phrase structures. As a typical example, there is an NE annotation that covers only part of a coordination structure, such as “Peter and Edward Bronfman”, where “Edward Bronfman” is annotated as an NE. In this case, we extend an original NE-span to the whole coordination structure.

We show a procedure to get an NE-aware dependency tree below.



Files

Phrase-structure

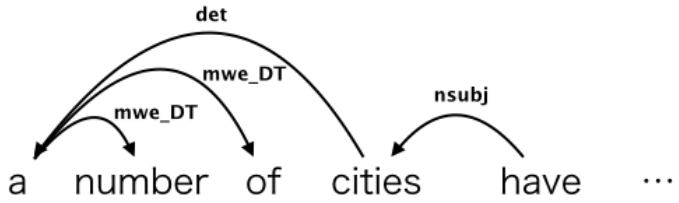
phrase_structures/[section]/*.parse

- MWE-aware phrase structure trees. As a non-terminal of a subtree corresponding to an MWE, we use NP_NE_<TYPE> or MWE_<POS>. The former (e.g. NP_NE_PERSON) is for named entities, and the latter (e.g. MWE_DT) is for compound function words.

Dependency

dependency/ontonotes_wsj_00_24_ne_mwe_aware_head_initial.conll

- Head-initial dependency structures which encode MWE-spans and MWE-level POS tags.



Conll Format

- 1 token per line, with blank lines separating sentences.
- 9 tab-separated columns (columns 1-8 are based on CoNLL-X Format [6]):
  1. ID
  2. FORM
  3. LEMMA
  4. CPOSTAG (filled by underscore)
  5. POSTAG
  6. FEATS (filled by underscore)
  7. HEAD
  8. DEPREL
  9. Filename in Ontonotes (e.g. wsj_0001)

References

- [1] Akihiko Kato, Hiroyuki Shindo and Yuji Matsumoto. 2017. English Multiword Expression-aware Dependency Parsing including Named Entities. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (To Appear).
- [2] Akihiko Kato, Hiroyuki Shindo and Yuji Matsumoto. 2016. Construction of an English Dependency Corpus incorporating Compound Function Words. Proceedings of 10th edition of the Language Resources and Evaluation Conference, pages 1667-1671, Portorož, Slovenia. (http://www.lrec-conf.org/proceedings/lrec2016/pdf/422_Paper.pdf)
- [3] Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, Yuji Matsumoto. 2013. Construction of English MWE Dictionary and its Application to POS Tagging. Proceedings of the 9th Workshop on Multiword Expressions, pages 139–144, Atlanta, Georgia, USA. Association for Computational Linguistics. (http://www.aclweb.org/anthology/W13-1021)
- [4] Marie-Catherine de Marneffe, Christopher D. Manning. 2008. The Stanford Typed Dependencies Representation. Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee. (http://www.aclweb.org/anthology/W08-1301)

- [5] Ryan Mcdonald, Joakim Nivre, Yvonne Quirmbach- brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Ta ̈ckstro ̈m, Claudia Bedini, Nu ́ria Bertomeu Castello ́, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 92–97. (https://aclweb.org/anthology/P/P13/P13-2017.pdf)
- [6] CoNLL-X Shared Task: Multi-lingual Dependency Parsing (http://ilk.uvt.nl/conll/)

History

- MWE-aware Dependency 1.0: 2015-10-23.
- MWE-aware Dependency 1.1: 2016-06-07. We revised statements about license.
- MWE-aware Dependency 2.0: 2017-07-21. We extended target MWEs to named entities.

Contact

- Please e-mail kato.akihiko.ju6 /at/ is.naist.jp with questions.

Contributors

- Akihiko Kato
- Hiroyuki Shindo
- Yuji Matsumoto