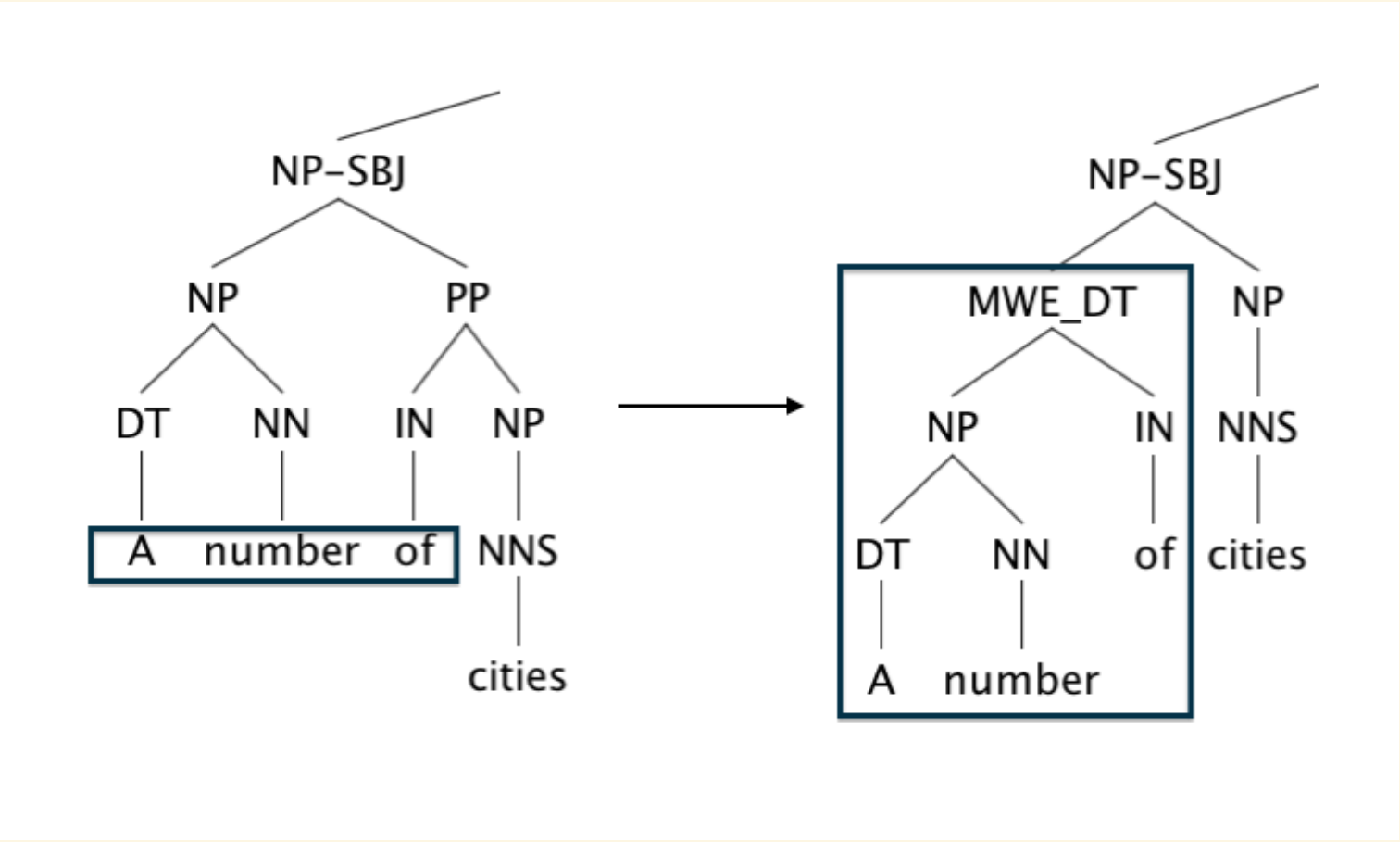
MWE-aware English Dependency Corpus

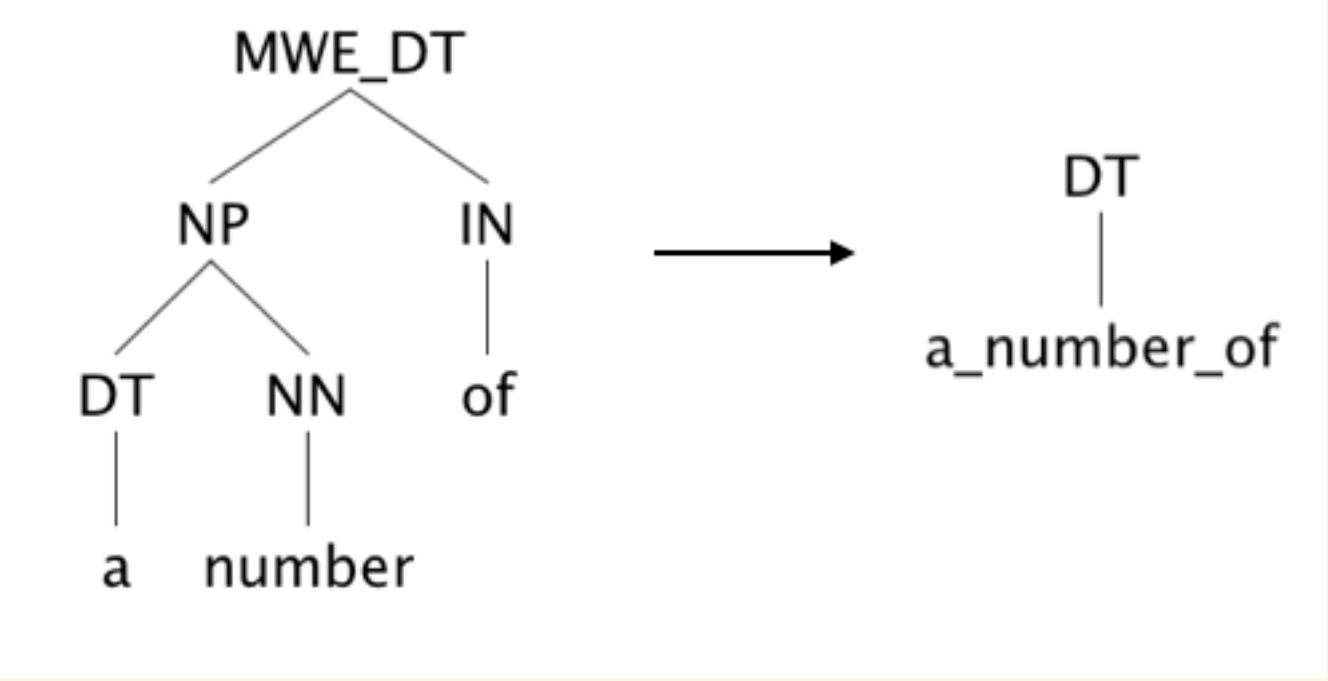
We provide users with an English dependency corpus taking into account compound function words, which are one type of multiword expressions (MWEs) and serve as functional expressions.

We built the corpus according to the following method [1].

1. We found an MWE in the phrase structure trees of Ontonotes and establish it as a single subtree.
    - We utilized the information (position in sentence and part of speech) of MWEs provided by [2].
    - The phrase structure trees made by this step are also provided.

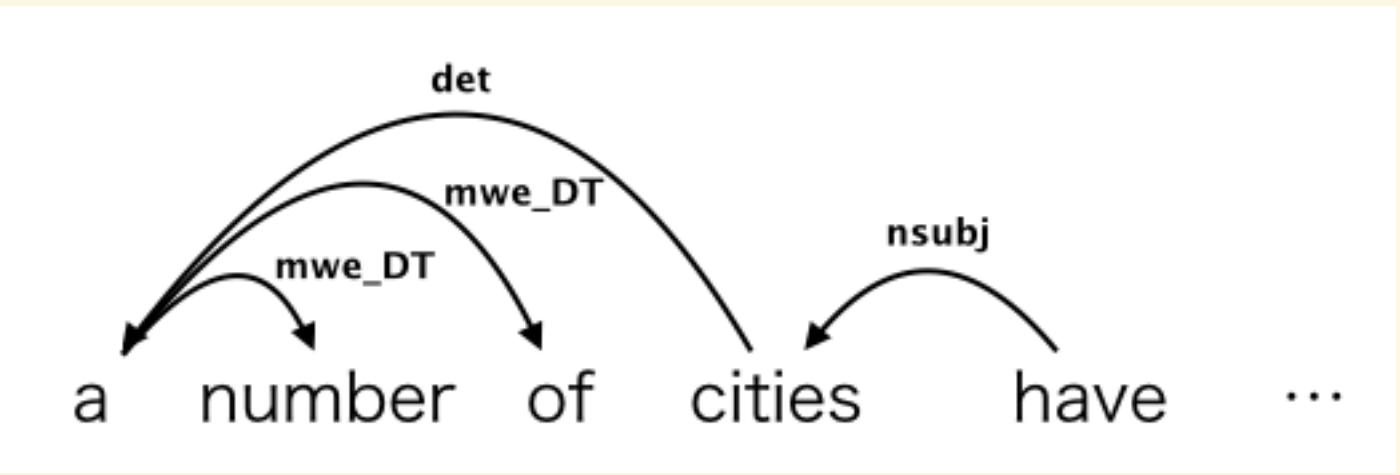


2. We replaced the above subtree by a preterminal with its leaf node as a child. The preterminal has the same part of speech as that of the MWE. Its child node is made by joining all components of the MWE with underscores.



3. We converted the phrase structure into Stanford Dependency [3].
    - We designated "-conllx -basic -makeCopulaHead -keepPunct" as a option for the conversion command.
    - We show an example of MWE-aware Dependency here.

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



Note

This language resource is distributed by LDC, because this corpus is based on Ontonotes release-5.0 (LDC2013T19).

Files

Phrase-structure (phrase_structure/ontonotes_5.0_mwe_aware_v1.0/wsj)

- MWE-aware phrase structure trees based on the Wall Street Journal portion in Ontonotes release-5.0 (LDC2013T19).

Dependency (dependency/ontonotes_wsj_00_24_mwe_aware.conll)

- MWE-aware Dependency for Section 00-24 of Wall Street Journal in Ontonotes (Stanford Dependency).

Conll Format

1 token per line, with blank lines separating sentences.

14 tab-separated columns (columns 1-10 are based on CoNLL-X Format [5]):

1. ID
2. FORM
3. LEMMA
4. CPOSTAG (filled by underscore)
5. POSTAG
6. FEATS (filled by underscore)
7. HEAD
8. DEPREL
9. PHEAD (filled by underscore)
10. PDEPREL (filled by underscore)
11. Filename in Ontonotes (e.g. wsj_0001)

12. Head in MWE_aware dependency (for MWEs, we adopt "head-initial" structure.)
13. Dependency label in MWE_aware dependency (we use "mwe" label for each token of MWE excluding the first token)
14. Part-of-speech tag of MWE (only for the first token of each MWE)

References

- [1] 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)

- [2] 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)

- [3] 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)

- [4] 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)

- [5] 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. (revised statements about license)

Contact

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

Contributors

- Akihiko Kato
- Hiroyuki Shindo
- Yuji Matsumoto