Cooperating Tools for MWE Lexicon Management and Corpus Annotation

Yuji Matsumoto
Nara Institute of Science & Technology
matsu@is.naist.jp

Akihiko Kato
Nara Institute of Science & Technology
kato.akihiko.ju6@is.naist.jp

Hiroyuki Shindo
Nara Institute of Science & Technology
shindo@is.naist.jp

Toshio Morita
Sowa Giken
morita@sowa.com

Abstract

We present tools for lexicon and corpus management that offer cooperating functionality in corpus annotation. The former, named Cradle, stores a set of words and expressions where multi-word expressions are defined with their own part-of-speech information and internal syntactic structures. The latter, named ChaKi, manages text corpora with part-of-speech (POS) and syntactic dependency structure annotations. Those two tools cooperate so that the words and multi-word expressions stored in Cradle are directly referred to by ChaKi in conducting corpus annotation, and the words and expressions annotated in ChaKi can be output as a list of lexical entities that are to be stored in Cradle.

1 Introduction

This paper presents tools for corpus and lexicon management, especially based on syntactic dependency structures. Annotating multi-word expressions (MWEs) in POS-tagged and/or syntactically analyzed corpora pose a number of problems. Out of three dependency relations for MWE annotations defined in Universal Dependency (UD) (Nivre et al., 2016) [1], only the `compound` relation can define internal syntactic structures, whereas the other two, `fixed` and `flat`, annotate MWEs with flat structures. Some issues of MWE annotation in UD are discussed in (Kahane et al., 2017). Since some MWEs have clear internal structure and they themselves interact with other words in the sentence with their own syntactic functionality, representing all their information solely in a dependency structure is not easy, or even possible. An MWE may be included in another flexible MWE. While an extended hierarchical BIO annotation scheme was proposed in (Schneider et al, 2002), it still cannot represent nested interaction of more than two MWEs.

In this paper, we do not discuss these issues in detail since the annotation standards are changing and there can be several standards for MWE annotation. We rather introduce two annotation tools that we are developing that are adaptable to various annotation standards as far as they are based on dependency syntax. The examples we show in this paper adopt the Stanford Typed Dependency (de Marneffe et al., 2008). One tool is a dictionary management tool named Cradle, which allows for the representation of the internal structures of MWEs, and stores them as a lexical resource. The other is an annotated corpus management tool called ChaKi, which communicates with Cradle and uses the stored lexical information during the corpus annotation process. While both tools are language independent and are currently used to develop English, Japanese, and Chinese dictionaries and corpora, we use English to explain how these tools cooperate in handling MWEs in corpus annotation.

2 Dictionary Management Tool: Cradle

The tool called Cradle was developed to maintain multi-lingual dictionaries. It stores words (including multi-word expressions), their part-of-speech labels, and additional information (such as inflection types, lemma forms, and internal structures for MWEs). Other than the information about individual

[1] Universal Dependency version 2: http://universaldependencies.org/

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 201–206
Santa Fe, New Mexico, USA, August 25-26, 2018.

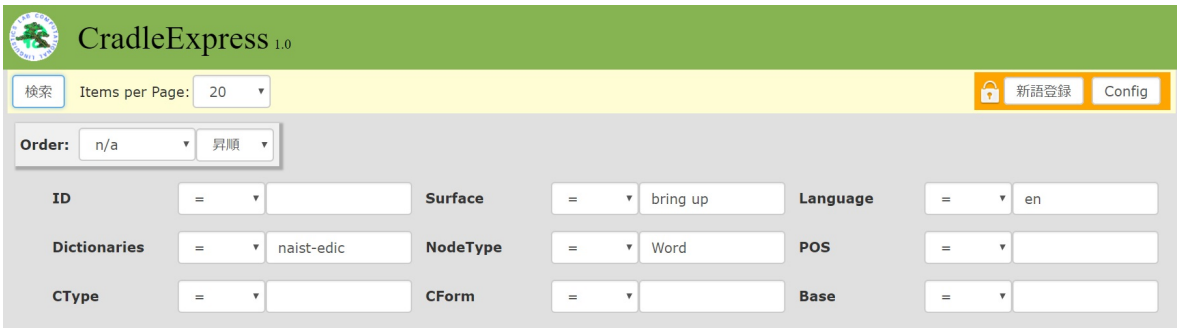

Figure 1: Snapshot of Cradle's start page

words/expressions, it retains the derivation relation between a base word and its derived forms (for now only inflected forms of verbs, nouns, adjectives, and adverbs) and the translation relation between pairs of translated words in two or more languages.

In this paper, we focus in particular on how MWEs are represented in the system. MWEs are categorized into fixed expressions, semi-fixed expressions, syntactically-flexible expressions, and institutional phrases (Sag et al, 2002). While all those types of MWEs can be handled in Cradle, we mainly focus on the first three types of English MWEs in this paper.

Figure 1 shows a snapshot of the start page of Cradle. When retrieving words or expressions, the user can optionally specify the language and the name of the dictionary. MWEs are defined as entries that include more than one word. Regular expressions can be used for retrieving words/expressions. In Figure 1, POS means part-of-speech and CType means the type of conjugation (or inflection). In the case of English, it is either "regular" or "irregular" and is only defined for verbs, nouns, adjectives, and adverbs. Figure 2 shows the word information pane that shows the basic information of the retrieved word (in this case "bring up"), its part-of-speech (POS), conjugation type (CType), and lemma form (Base). In the case of MWEs, we can specify not only their basic information like POS, CType, and lemma form, but also their internal dependency structures.

Key	Value
ID	526a693414cb1719f48a5177
Surface	bring up
Language	en
Dictionaries	
NodeType	Word
DummyKind	someone,something
POS	VB-VB
CType	irregular
CForm	
Base	bring up

Figure 2: Word information

Examples of fixed multi-word expressions are "with respect to" and "lots of," which behave as single lexical entries with their respective parts-of-speech. The former behaves as a preposition and the latter behaves as a determiner as a whole. Semi-fixed and syntactically-flexible expressions may have modifiers within them or allow for syntactic variations. For example, while "a number of" behaves as a determiner, it has some variations such as "a small number of" or "a large number of," and both behave as a determiner with additional meaning. Another flexible expression "take into account" behaves as a transitive verb associated with a direct object, but is often used as "take *something* into account," taking a direct object within the expression. Figure 3 explains how the dependency structures of such MWEs are defined in Cradle. In the figure, an asterisk defines a "placeholder" that matches any word. While it is not visible on this screen, the POS tag information can be specified on the placeholder as a constraint. In the case of "a $*$ number of $*$," the first placeholder stands for an adjective and the second stands for a noun. In the case of "take $*$ into account," the placeholder stands for a noun. The rightmost box without any word or asterisk is a dummy box for pointing to the head word. The reason that the words are depicted as sequences of characters is for handling non-segmented languages like Japanese and Chinese, in which annotators need to segment a sequence of characters into words when they first define the structure of MWEs.

In our English dictionary, we define an extended version of the POS tag set of PennTreebank (Marcus et al., 1993). All the POS tags are define by two tag layers. For example, the POS tags of singular and

plural nouns “NN” and “NNS” are defined as “NN-NN” and “NN-NNS,” respectively. Similarly, “VBP” and “VBD” are defined as “VB-VBP” and “VB-VBD” so that the second layer corresponds exactly to the PTB POS tags and the first layer defines the coarse-grained tag groups. An adjective, a noun, and a verb (in any form) are defined as “JJ-∗,” “NN-∗,” and “VB-∗.” Moreover, “a ∗ number of ∗” and “take ∗ into account” are defined as a determiner and a verb for their own parts-of-speech.

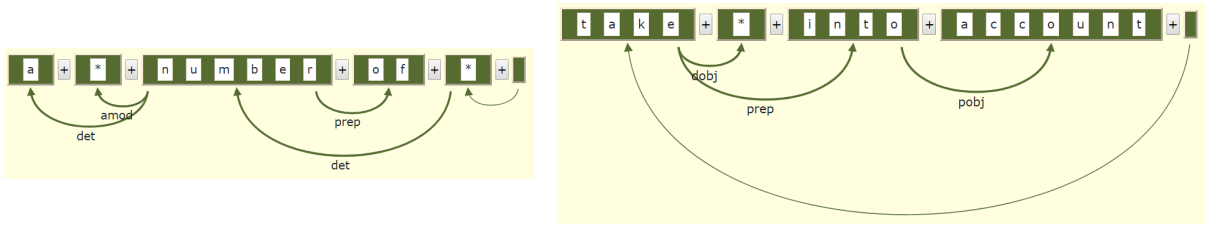

Figure 3: Dependency annotation of MWEs in Cradle

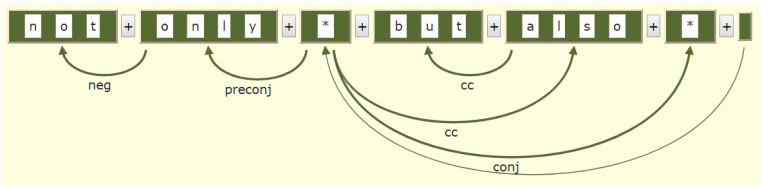

Figure 4: Dependency annotation of “not only ··· but also ···”

There are some MWEs that do not have an appropriate POS tag of their own. Our English dictionary includes flexible MWEs that do not stand for a single POS tag but define a complex syntactic structure. One such example is “not only ··· but also ···,” which constitutes a coordination. Figure 4 shows the dependency structure associated with this expression. The two placeholders are the heads of the conjuncts connected by this MWE. While they should share a common syntactic property such as a common POS, the current Cradle does not provide such a functionality (equivalence relation between POS tags).

3 Corpus Management Tool: ChaKi

This section introduces a tool called ChaKi, which helps to annotate corpora with POS and dependency information. It also offers flexible search operations for sentences using character sequences, words with various lexical properties, and dependency relations. Other than POS and word dependency annotation, ChaKi is equipped with segment and group annotation functions. A sequence of words are put together in a segment, and two or more segments are related in a group. A segment can be used to define named entities or fixed/flat multi-word expressions. Two or more segments are related to form a group.

Figure 5 shows a snapshot of ChaKi when a user retrieves sentences that include the expression “but also.” The top box shows the pattern of the sentence retrieval, and the middle window shows three sentences that include this expression. The lowest window, the dependency pane, shows a part of the dependency tree of the first sentence. There are two groups as shown in this sentence: The first corresponds to “not only ··· but also ···” (green boxes labeled by “1:MWE”), and the second corresponds to the coordination structure conjoined by this MWE (orange boxes labeled with “2:Parallel”). Segment and group annotation as well as other annotation such as the POS tag and dependency edge correction/modification can be done via mouse operations on this interface.

When we annotate MWEs that have specific syntactic functions like determiners or adverbs, it is not only their internal dependency structure but also their dependency relation with other words in the sentence that may need to be modified. For example, Figure 6 shows how the dependency annotation should be changed after an expression is found as an MWE. The left-hand side shows the original dependency tree. Note that the syntactic head of the noun phrase “a number of southern states” is “number” in this tree. When “a number of” in this sentence is found as an MWE that functions as a determiner, they are to

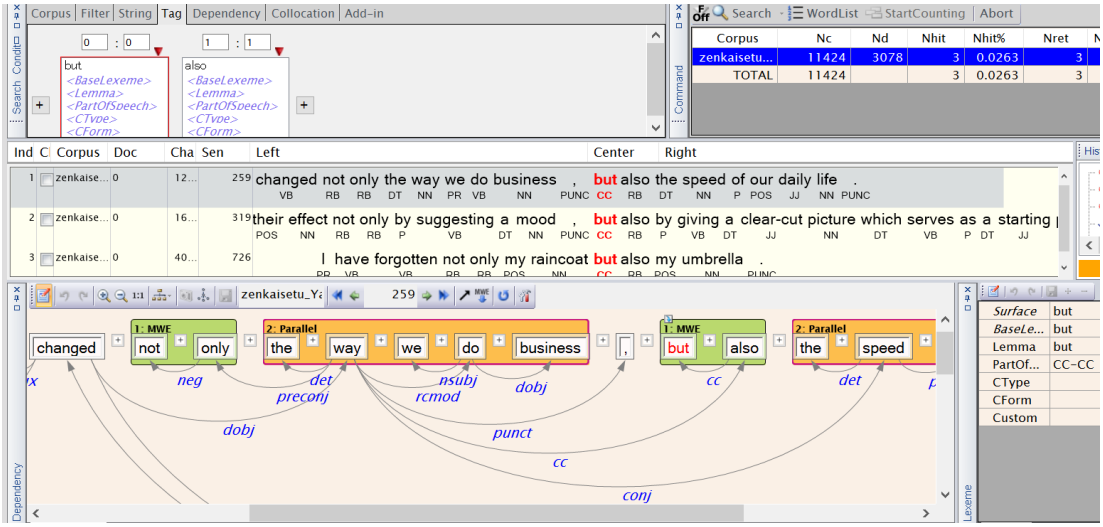

Figure 5: Snapshot of ChaKi in retrieving “but also”

be annotated in a group (shown by a green box) as shown on the right-hand side. Simultaneously, since it is regarded as a single determiner, the head of the noun phrase should be “states” and the dependency relations need to be corrected as shown in the figure.

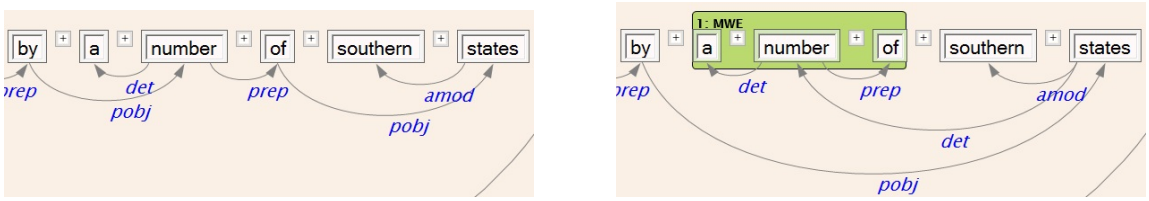

Figure 6: Bare annotation and MWE-aware annotation of “a number of”

While ChaKi offers intuitive mouse operations to annotate segments and groups and to modify dependency edges, it is very tiresome to do the same operation when the same MWEs appear repeatedly in a corpus. This is why we implemented the close cooperation of the above two tools, which is described in the next section.

4 Cooperation between Cradle and ChaKi

To maintain the internal structures and other lexical information of MWEs and to utilize them for efficient corpus annotation, we implemented cooperation functions that connect Cradle and ChaKi. ChaKi can refer to the information in Cradle in the annotation process. Suppose an annotator is working on POS and dependency structure annotation or on the correction of the following sentence:

> “Currently, a great number of electric cars are operating on Japanese streets.”

By pressing the MWE button, ChaKi presents a list of all possible MWEs and their positions in the sentence by consulting the dictionary in Cradle; the MWEs and their positions are shown in Figure 7. There appears a flexible MWE, “a great number of,” in this sentence, and the list shows all possible matching positions of this MWE. Note that it is defined as “a (JJ-∗) number of (NN-∗)” in Cradle. Each line in Figure 7 corresponds to an MWE, its position. For example, the second line indicates that “a (great) number of (cars)” matches with an MWE. The column “WordPositions” shows the word positions of the MWE in the sentence, meaning that the position of “a” is at 2, and that of “number” is at 4, etc. When a user selects a line by clicking the button on the “Apply” column, the dependency structure of the

MWE is tentatively shown on the dependency pane. The user then presses the apply button (not shown on the figure) to effect the annotation.

	Apply	Surface	WordPositions
▶	☐	a number of	2,4,5
	☐	a (great) number of (streets)	2,3,4,5,12
	☐	a (great) number of (cars)	2,3,4,5,7

Figure 7: List of possible MWEs included in the current sentence

The left-hand side of Figure 8 shows the original dependency structure of the sentence before MWE annotation. After checking the MWE list shown in Figure 7 and applying the second line, the annotation on the right-hand side of Figure 8 is obtained automatically.

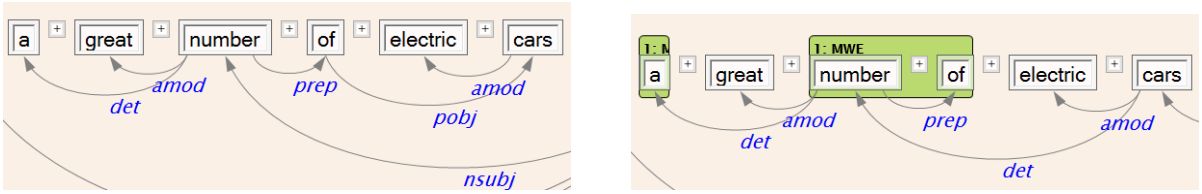

Figure 8: Before and after MWE annotation of "a great number of"

Currently, ChaKi can consult the dictionary defined in Cradle and conduct efficient annotation of MWEs (as well as other lexical entries) as explained above. If the user does not find any definition of an MWE in Cradle, he/she can group the words together to form an MWE group, modify the dependency structure, and output the result. The output format is based on the CoNLL-U format [2] with some additional information for MWE groups and their POS tags. The dictionary manager needs to examine which MWEs are to be registered in Cradle, and may need to add additional information like POS constrains on placeholders and modify the dependency structures if necessary.

5 Conclusion

This paper presented our tools for managing dictionaries and (POS and dependency) annotated corpora. The system for dictionary management, Cradle, stores words and multi-word expressions together with their POS and syntactic dependency structure information. The system for corpus annotation, ChaKi, deals with POS and dependency annotated corpora to retrieve sentences and to modify or correct annotations including MWE annotation. We presented the main functions of those two tools and introduced the cooperation between them that is especially effective for MWE annotation by simply selecting possible MWEs defined in the dictionary.

There are several dependency annotation schemes in many languages, such as Universal dependency and CoNLL dependency. Initially, our dependency definition adopted the CoNLL dependency and then the Stanford dependency. Since MWEs often cause problematic cases of syntactic annotation, our framework of managing MWEs and their syntactic information in a dictionary provides an easy and consistent way of coping with definition changes or transferring one scheme to another.

Currently, Cradle is not publicly available, but is available on request. ChaKi is available from the following site:

https://ja.osdn.net/projects/chaki/

Acknowledgements

This work was partially supported by JST CREST Grant Number JPMJCR1513, Japan.

[2] http://universaldependencies.org/format.html

References

Marie-Catherine de Marneffe and Manning, C.D. 2008. Stanford Typed Dependencies Manual (revised in 2016). Stanford University.

Marie-Catherine de Marneffe, M-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C. 2014. Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4585–4592, Reykjavik, Iceland.

Kahane, S., Courtin, M. and Gerdes, K. 2017. Multi-word annotation in Syntactic treebanks: Proposition for Universal Dependency. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 181–189, Prague, Czech Republic.

Marcus, M. et al. 1993. Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19(2):313–330.

Nivre, J., de Marneffe, M-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Sag, I., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing*, Springer-Verlag, 1–15.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M., Conrad, H. and Smith, N. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, 455–461, Reykjavik, Iceland.