Semi-Automated Resolution of Inconsistency for a Harmonized Multiword Expression and Dependency Parse Annotation

King Chan, Julian Brooke, and Timothy Baldwin

Department of Computing and Information Systems, The University of Melbourne

chanking@gmail.com, julian.brooke@unimelb.edu.au, tb@ldwin.net

Abstract

This paper presents a methodology for identifying and resolving various kinds of inconsistency in the context of merging dependency and multiword expression (MWE) annotations, to generate a dependency treebank with comprehensive MWE annotations. Candidates for correction are identified using a variety of heuristics, including an entirely novel one which identifies violations of MWE constituency in the dependency tree, and resolved by arbitration with minimal human intervention. Using this technique, we identified and corrected several hundred errors across both parse and MWE annotations, representing changes to a significant percentage (well over 10%) of the MWE instances in the joint corpus.

1 Introduction

The availability of gold-standard annotations is important for the training and evaluation of a wide variety of NLP tasks, including the evaluation of dependency parsers (Buchholz and Marsi, 2006). In recent years, there has been a focus on multi-annotation of a single corpus, such as joint syntactic, semantic role, named entity, coreference and word sense annotation in Ontonotes (Hovy et al., 2006) or constituency, semantic role, discourse, opinion, temporal, event and coreference (among others) annotation of the Manually Annotated Sub-Corpus of the ANC (Ide et al., 2010). As part of this, there has been an increased focus on harmonizing and merging existing annotated data sets as a means of extending the scope of reference corpora (Ide and Suderman, 2007; Declerck, 2008; Simi et al., 2015). This effort sometimes presents an opportunity to fix conflicting annotations, a worthwhile endeavour since even a small number of errors in a gold-standard syntactic annotation can, for example, result in significant changes in downstream applications (Habash et al., 2007). This paper presents the results of a harmonization effort for the overlapping STREUSLE annotation (Schneider et al., 2014) of multiword expressions ("MWEs": Baldwin and Kim (2010)) and dependency parse structure in the English Web Treebank ("EWT": Bies et al. (2012)), with the long-term goal of building reliable resources for joint MWE/syntactic parsing (Constant and Nivre, 2016).

As part of merging these two sets of annotations, we use analysis of cross-annotation and type-level consistency to identify instances of potential annotation inconsistency, with an eye to improving the quality of the component and combined annotations. It is important to point out that our approach to identifying and handling inconsistencies does not involve re-annotating the corpus; instead we act as arbitrators, resolving inconsistency in only those cases where human intervention is necessary. Our three methods for identifying potentially problematic annotations are:

- a cross-annotation heuristic that identifies MWE tokens whose parse structure is incompatible with the syntactic annotation of the MWE;
- a cross-type heuristic that identifies n-grams with inconsistent token-level MWE annotations; and
- a cross-type, cross-annotation heuristic that identifies MWE types whose parse structure is inconsistent across its token occurrences.

The first of these is specific to this harmonization process, and as far as we aware, entirely novel. The other two are adaptions of an approach to improving syntactic annotations proposed by Dickinson and Meurers (2003). After applying these heuristics and reviewing the candidates, we identified hundreds of errors in MWE annotation and

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 187–193,
Valencia, Spain, April 4.

about a hundred errors in the original syntactic annotations. We make available a tool that applies these fixes in the process of joining the two annotations into a single harmonized, corrected annotation, and release the harmonized annotations in the form of HAMSTER (the HArmonized Multiword and Syntactic TreE Resource): `https://github.com/eltimster/HAMSTER`.

2 Related Work

Our long-term goal is in building reliable resources for joint MWE/syntactic parsing. Explicit modelling of MWEs has been shown to improve parser accuracy (Nivre and Nilsson, 2004; Finkel and Manning, 2009; Korkontzelos and Manandhar, 2010; Green et al., 2013; Vincze et al., 2013; Candito and Constant, 2014; Constant and Nivre, 2016). Treatment of MWEs has typically involved parsing MWEs as single lexical units (Nivre and Nilsson, 2004; Eryiğit et al., 2011; Aggeliki Fotopoulou, 2014), however this flattened, "words with spaces" (Sag et al., 2002) approach is inflexible in its coverage of MWEs where components have some level of flexibility.

The English Web Treebank (Bies et al., 2012) represents a gold-standard annotation effort over informal web text. The original syntactic constituency annotation of the corpus was based on hand-correcting the output of the Stanford Parser (Manning et al., 2014); for our purposes we have converted this into a dependency parse using the Stanford Typed Dependency converter (de Marneffe et al., 2006). We considered the use of the Universal Dependencies representation (Nivre et al., 2016), however we noted that several aspects of that annotation (in particular the treatment of all prepositions as case markers dependent on their noun) make it inappropriate for joint MWE/syntactic parsing since it results in large numbers of MWEs that are non-contiguous in their syntactic structure (despite being contiguous at the token-level). As such, the Stanford Typed Dependencies are the representation which has the greatest currency for joint MWE/syntactic parsing work (Constant and Nivre, 2016).

The STREUSLE corpus (Schneider et al., 2014) is based entirely on the Reviews subset of the EWT, and comprises of 3,812 sentences representing 55,579 tokens. The annotation was completed by six linguists who were native English speakers. Every sentence was assessed by at least two annotators, which resulted in an average inter-annotator F1 agreement of 0.7. The idiosyncratic nature of MWEs lends itself to challenges associated with their interpretation, and this was readily acknowledged by those involved in the development of the STREUSLE corpus (Hollenstein et al., 2016). Two important aspects of the MWE annotation are that it includes both contiguous and non-contiguous MWEs (e.g. *check* $*$ *out*), and that it supports both weak and strong annotation; both of these are considered in scope for our inconsistency analysis. A variety of cues are employed to determine this associative strength. The primary factor relates to the degree in which the expression is semantically opaque and/or morphosyntactically idiosyncratic. An example of a strong MWE would be *top notch*, as used in the sentence: *We stayed at a top notch hotel.* The semantics of this expression are not immediately predictable from the meanings of *top* and *notch*. On the other hand, the expression *highly recommend* is considered to be a weak expression as it is largely compositional — one can *highly recommend a product* — as indicated by the presence of alternatives such as *greatly recommend* which are also acceptable though less idiomatic. A total of 3,626 MWE instances were identified in STREUSLE, across 2,334 MWE types.

Other MWE-aware dependency treebanks include the various UD treebanks (Nivre et al., 2016), the Prague Dependency Treebank (Bejček et al., 2013), and others (Nivre and Nilsson, 2004; Eryiğit et al., 2011; Candito and Constant, 2014). The representation of MWEs, and the scope of types covered by these treebanks, can vary significantly. For example, the internal syntactic structure may be flattened (Nivre and Nilsson, 2004), or in the case of Candito and Constant (2014), allow for distinctions in the granularity of syntactic representation for regular vs. irregular MWE types.

The identification of inconsistencies in annotation requires comparisons to be made between similar instances that are labeled differently. Boyd et al. (2007) employed an alignment-based approach to assess differences in the annotation of n-gram word sequences in order to establish the likelihood of error occurrence. Other work in the syntactic inconsistency detection domain includes those related to POS tagging (Loftsson, 2009; Eskin, 2000; Ma et al., 2001) and parse structure (Ule and Simov, 2004; Kato and Mat-

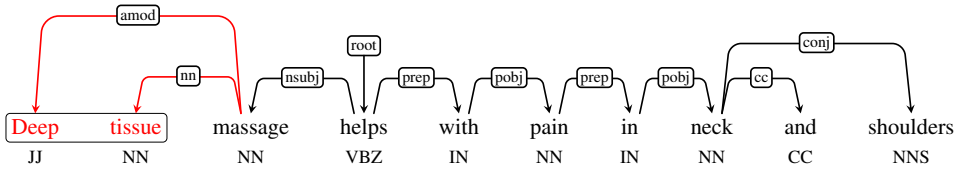

Figure 1: An example where the arc count heuristic is breached. *Deep tissue* has been labeled in the sentence here as an MWE in STREUSLE. *Deep* and *tissue* act as modifiers to *massage*, a term that has not been included as part of the MWE.

subara, 2010). Dickinson and Meurers (2003) outline various approaches for detecting inconsistencies in parse structure within treebanks.

In general, inconsistencies associated with MWE annotation fall under two categories: (1) *annotator error* (i.e. false positives and false negatives); and (2) ambiguity associated with the assessment of *hard cases*. While annotation errors apply to situations where a correct label can be applied but is not done so, hard cases are those where the correct label is inherently difficult to assign, and can be particularly relevant to certain classes of MWEs. For example, there may be considerable differences in inter-annotator agreement associated with assessing the relative transparency and associative strength of a non-fixed MWE.

3 Error Candidate Identification

3.1 MWE Syntactic Constituency Conflicts

The hypothesis that drives our first analysis is that for nearly all MWE types, the component words of the MWE should be syntactically connected, which is to say that every word is a dependent of another word in the MWE, except one word which connects the MWE to the rest of the sentence (or the root of the sentence). We can realise this intuition by using an arc count heuristic: for each labeled MWE instance we count the number of incoming dependency arcs that are headed by a term outside the MWE, and if the count is greater than one, we flag it for manual analysis. Figure 1 gives an example where the arc count heuristic is breached since both terms of the MWE *deep tissue* act as modifiers to the head noun that sits outside the MWE.

3.2 MWE Type Inconsistency

Our second analysis involves first collecting a list of all MWE types in the STREUSLE corpus, corresponding to lemmatized n-grams, possibly with gaps. We then match these n-grams across the same corpus, and flag any MWE type which has at least one inconsistency with regards to the annotation. That is, we extract as candidates any MWE types where there were at least two occurrences of the corresponding n-gram in the corpus that were incompatible with respect to their annotation in STREUSLE, including discrepancies in weak/strong designation. For non-contiguous MWE types, matches containing up to 4 words of intervening context between the two parts of the MWE type were included as candidates for further assessment.

3.3 MWE Type Parse Inconsistency

The hypothesis that drives our third analysis is that we would generally expect the internal syntax of an MWE type to be consistent across all its instances.[1] For each MWE type, we extracted the internal dependency structure of all its labeled instances, and flagged for further assessment any type for which the parse structure varied between at least two of those instances. Note that although this analysis is aimed at fixing parse errors, it makes direct use of the MWE annotation provided by STREUSLE to greatly limit the scope of error candidates to those which are most relevant to our interest.

4 Error Arbitration

Error arbitration was carried out by the authors (all native English speakers with experience in MWE identification), with at least two authors looking at each error candidate in most instances, and for certain difficult cases, the final annotation being based on discussion among all three authors. One advantage of our arbitration approach over a traditional token-based annotation was that we could enforce consistency across similar error can-

[1] Noting that we would not expect this to occur between MWE instances of a given combination of words, and non-MWE combinations of those same words.

didates (e.g. *disappointed with* and *happy with*) and also investigate non-candidates to arrive at a consensus; where at all possible, our changes relied on precedents that already existed in the relevant annotation.

Arbitration for the MWE syntax conflicts usually involved identifying an error in one of the two annotations, and in most cases this was relatively obvious. For instance, in the candidate ... *the usual lady called in sick hours earlier*, *called in sick* was correctly labeled as an MWE, but the parse incorrectly includes *sick* as a dependent of *hours*, rather than *called in*. An example of the opposite case is ... *just to make the appointment* ..., where *make the* had been labeled as an MWE, an obvious error which was caught by our arc count heuristic. There were cases where our arc count heuristic was breached due to what we would view as a general inadequacy in the syntactic annotation, but we decided not to effect a change because the impact would be too far reaching; examples of this were certain discourse markers (e.g. *as soon as*), and infinitives (e.g. *have to complete* where the *to* is considered a dependent of its verb rather than of the other term in the MWE *have to*). The most interesting cases were a handful of non-contiguous MWEs where there was truly a discontinuity in the syntax between the two parts of the MWE, for instance *no amount of* $*$ *can*. This suggests a basic limitation in our heuristic, although the vast majority of MWEs did satisfy it.

For the two type-level arbitrations, there were cases of inconsistency upheld by real usage differences (e.g. *a little house* vs. *a little tired*). We identified clear differences in usage first, and divided the MWE types into sets, excluding from further analysis non-MWE usages of MWE type n-grams. For each consistent usage of an MWE type, the default position was to prefer the majority annotation across the set of instances, except when there were other candidates that were essentially equivalent: for instance, if we had relied on majority annotation for *job* $*$ *do* (e.g. *the job that he did*) it would have been a different annotation than *do* $*$ *job* (e.g. *do a good job*), so we considered these two together. We treated contiguous and non-contiguous versions of the same MWE type in the same manner.

In the MWE type consistency arbitration, for cases where majority rules did not provide a clear answer and there was no overwhelming evidence for non-compositionality, we introduced a special internal label called *hard*. These correspond to cases where the usage is consistent and the inconsistency seems to be a result of the difficulty of the annotation item (as discussed earlier in Section 2), which extended also to our arbitration. Rather than enforce a specific annotation without strong evidence, or allow the inconsistency to remain when there is no usage justification for it, the corpus merging and correction tool gives the user the option to treat *hard* annotated MWEs in varying ways: the annotation may be kept unchanged, removed, converted to weak, or covered to *hard* for the purpose of excluding it from evaluation. Examples of hard cases include *go back, go in, more than, talk to, speak to, thanks guys, not that great, pleased with, have* $*$ *option, get* $*$ *answer, fix* $*$ *problem*. On a per capita basis, inconsistencies are more common for non-contiguous MWEs relative to their contiguous counterparts, and we suspect that this is partially due to their tendency to be weaker, in addition to the challenges involved in correctly discerning the two parts, which are sometimes at a significant distance from each other.

Table 1 provides a summary of changes to MWE annotation at the MWE type and token levels. *Mixed* refer to MWEs that are heterogeneous in the associative strength between terms in the MWE (between `weak` and `strong`). Most of the changes in Table 1 (98% of the types) were the result of our type consistency analysis. Almost half of the changes involved the use of the `hard` label, but even excluding these (since only some of these annotations required actual changes in the final version of the corpus) our changes involve over 10% of the MWE tokens in the corpus, and thus represent a significant improvement to the STREUSLE annotation.

Relative to the changes to the MWE annotation, the changes to the parse annotation were more modest, but still not insignificant: for 181 MWE tokens across 157 types, we identified and corrected a dependency and/or POS annotation error. The majority of these (61%) were identified using the arc count heuristic. Note we applied the parse relevant heuristics after we fixed the MWE type consistency errors, ensuring that MWE annotations that were added were duly considered for parse errors.

		No MWE	Weak	Strong	Mixed	Hard	TOTAL
Token	**No MWE**	—	56	134	6	148	344
	Weak	33	—	22	5	46	106
	Strong	41	43	—	9	70	163
	Mixed	0	4	5	14	2	25
	TOTAL	74	103	161	34	266	638
Type	**No MWE**	—	31	72	5	63	171
	Weak	29	—	13	4	35	81
	Strong	32	28	—	7	43	110
	Mixed	0	4	4	9	2	19
	TOTAL	61	63	89	25	143	381

Table 1: Summary of changes to MWE annotation at the MWE type and token level

5 Discussion

Our three heuristics are useful because they identify potential errors with a high degree of precision. For the MWE type consistency analysis 77% of candidate types were problematic, and for parse type consistency, 79%. For the arc count heuristic, 45% of candidate types were ultimately changed: as mentioned earlier, some of the breaches involved systematic issues with annotation schema that we felt uncomfortable changing in isolation. By bringing these candidate instances to our attention, we were able to better focus our manual analysis effort, including in some cases looking across multiple related types, or even searching for specialist knowledge which could resolve ambiguities: for instance, in the example shown in Figure 1, though a layperson without reference material may be unsure whether it is *tissue* or *massage* which is considered to be *deep*, a quick online search indicates that the original EWT syntax is in error (*deep* modifies *tissue*).

However, it would be an overstatement to claim to have fixed all (or even almost all) the errors in the corpus. For instance, our type consistency heuristics only work when there are multiple instances of the same type, yet it is worth noting that 82% of the MWE types in the corpus are represented by a singleton instance. Our arc count heuristic can identify issues with singletons, but its scope is fairly limited. We cannot possibly identify missing annotations for types that were not annotated at least once. We might also miss certain kinds of systematic annotation errors, for instance those mentioned in De Smedt et al. (2015), though that work focused on the use of `mwe` dependency labels which are barely used in the EWT, one of the reasons a resource like STREUSLE is so useful.

6 Conclusion

We have proposed a methodology for merging multiword expression and dependency parse annotations, to generate HAMSTER: a gold-standard MWE-annotated dependency treebank with high consistency. The heuristics used to enforce consistency operate at the type- and cross-annotation level, and affected well over 10% of the MWEs in the new resource.

References

Voula Giouli Aggeliki Fotopoulou, Stella Markantonatou. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovskỳ, Anna Nedoluzhko, Jarmila Panevová, et al. 2013. Prague dependency treebank 3.0.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. technical report ldc2012t13. Technical report, Linguistic Data Consortium.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. 2007. Increasing the recall of corpus annotation error detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *The 52nd Annual Meeting of the Association for Computational Linguistics (ACl '14)*.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*, pages 161–171.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genova, Italy.

Koenraad De Smedt, Victoria Rosén, and Paul Meurer. 2015. Studying consistency in ud treebanks with iness-search. In *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 258–267.

Thierry Declerck. 2008. A framework for standardized syntactic annotation. In *Proceedings of the 2008 Language Resource and Evaluation Conference (LREC 08)*.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*.

Gülşen Eryiğit, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, SPMRL '11, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eleazar Eskin. 2000. Detecting errors within a corpus using anomaly detection. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 148–153, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 326–334, Stroudsburg, PA, USA. Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227, March.

Nizar Habash, Ryan Gabbard, Owen Rambow, Seth Kulick, and Mitchell P Marcus. 2007. Determining case in arabic: Learning complex linguistic behavior requires complex linguistic features. In *EMNLP-CoNLL*, pages 1084–1092.

Nora Hollenstein, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of the 2016 Langauage Resources and Evaluation Conference (LREC '16)*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 57–60, New York City, USA.

Nancy Ide and Keith Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. pages 68–73.

Yoshihide Kato and Shigeki Matsubara. 2010. Correcting errors in a treebank based on synchronous tree substitution grammar. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 74–79. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 636–644, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hrafn Loftsson. 2009. Correcting a pos-tagged corpus using three complementary methods. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–531. Association for Computational Linguistics.

Qing Ma, Bao-Liang Lu, Masaki Murata, Michnori Ichikawa, and Hitoshi Isahara. 2001. On-line error detection of annotated corpus using modular neural

networks. In *International Conference on Artificial Neural Networks*, pages 1185–1192. Springer.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, et al. 2016. Universal dependencies v1: A multilingual treebank collection.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '02)*.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 455–461, Reykjavík, Iceland.

Maria Simi, Simonetta Montemagni, and Cristina Bosco. 2015. Harmonizing and merging italian treebanks: Towards a merged italian dependency treebank and beyond. In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 3–23.

Tylman Ule and Kiril Simov. 2004. Unexpected productions may well be errors. In *Proceedings of the 2004 Language Resources and Evaluation Conference (LREC '04)*.

Veronika Vincze, Janos Zsibrita, and István Nagy T. 2013. Dependency parsing for identifying hungarian light verb constructions. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP '13)*.