Rollenwechsel-English: a large-scale semantic role corpus

Asad Sayeed[1], **Pavel Shkadzko**[2], **Vera Demberg**[2]
[1]CLASP, Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg
[2]Saarland University
asad.sayeed@gu.se, p.shkadzko@gmail.com, vera@coli.uni-saarland.de

Abstract

We present the Rollenwechsel-English (RW-eng) corpus, a large corpus of automatically-labelled semantic frames extracted from the ukWaC corpus and BNC using Propbank roles. RW-eng contains both full-phrase constituents for labelled roles as well as heads identified by a series of heuristics. This corpus is of a scale and size suitable for new deep learning approaches to language modelling and distributional semantics, particularly as it pertains to generalized event knowledge. We describe the structure of this corpus, tools for its use, and successful use cases.

Keywords: semantic roles, web corpus, labelled data

1. Motivation

Semantic role labelling is a comparatively mature task in natural language processing. Typically, it takes the form of supervised classification or language modeling task, although there is more recent work in unsupervised induction of semantic roles (Titov and Klementiev, 2012). These approaches tend to have application goals in traditional areas of text-based natural language processing.

A major area of psycholinguistic research is the influence of semantic structure on adult sentence processing. Does the association between "cake" and "cutting" have an effect on processing difficulty of future sentence constituents in a sentence that starts with "The child cut the cake…", and to what extent is this association influenced by the semantic role-based selectional preferences of the verb (McRae et al., 1998; McRae et al., 2005; Ferretti et al., 2001; Bicknell et al., 2010)? A further question pertains to the internal representation of linguistic knowledge: to what extent do distributions of roles observed in text really reflect the internal state of human knowledge (generalized event knowledge)? Human being easily conceive of a knife as a proper instrument for cake-cutting—but knowledge of affordances allows them also to see other sharp objects (e.g. swords, floss) as potential cake-cutters, even if the co-occurrence frequency in the corpus is low. This possible mismatch between corpus frequency and the underlying cognitive model (Amsel et al., 2015) affects such application areas as dialogue systems or indeed any application in which intuitions similar to human ones are required.

Nevertheless, the expansion of corpus availability, increase in computing power, and powerful extension of traditional machine learning techniques such as deep learning provides new opportunities to understand these questions. What has been missing until recently, however, is a variety of data sources capable of supporting the type of hypothesis-testing about the ability of the new techniques to acquire the latent information about semantic relationships within the sentence.

We contribute towards addressing this in this paper by introducing the Rollenwechsel-English (RW-eng) corpus. RW-eng[1] is labelled automatically with semantic roles which are then reprocessed heuristically to yield a rich representation of verb-noun relationships and subcategorization frames. RW-eng is based on the full ukWaC corpus (Ferraresi et al., 2008) and the British National Corpus (BNC Consortium, 2007). The semantic role labelling is done by SENNA (Collobert and Weston, 2007; Collobert et al., 2011), a labeller that does not, as most other SRL tools do, rely directly on the syntactic parse of the sentence, allowing it to capture relationships that syntax-based SRL does not, and therefore implicitly permitting some investigation of semantic roles that are not totally confounded with an underlying syntactic theory.

In the remainder of this paper, we outline the process by which the sentences are labelled, the output format of the corpus and thus the data and relationships the corpus encodes, and publicly-available tools for corpus generation and access. We also discuss some scientific use cases for the corpus and results already obtained from it.

2. Corpus generation

2.1. Preprocessing

The bulk of our Python-based processing pipeline (`ukwac2tensor`) for the RW-eng corpus is available as a `git` repository on the web[2]. The initial input for the pipeline are the ukWaC and BNC corpora parsed by MaltParser (Nivre et al., 2007) and tagged with part-of-speech labels, in a column format supplied to us by the creators of ukWaC. For parallelization purposes, we divide the corpus up into 3500 segments approximately equal by number of documents. This division stays with the pipeline to the final output.

Since the ukWaC data comes from web data that is sometimes not codepage-consistent, we replaced or removed special characters so that it was compliant with UTF-8. Our goal was to produce XML output that would validate properly with common Python-based XML parsers.

[1] `http://rollen.mmci.uni-saarland.de/RW-eng/`

[2] `https://github.com/tastyminerals/ukwac2tensor/`

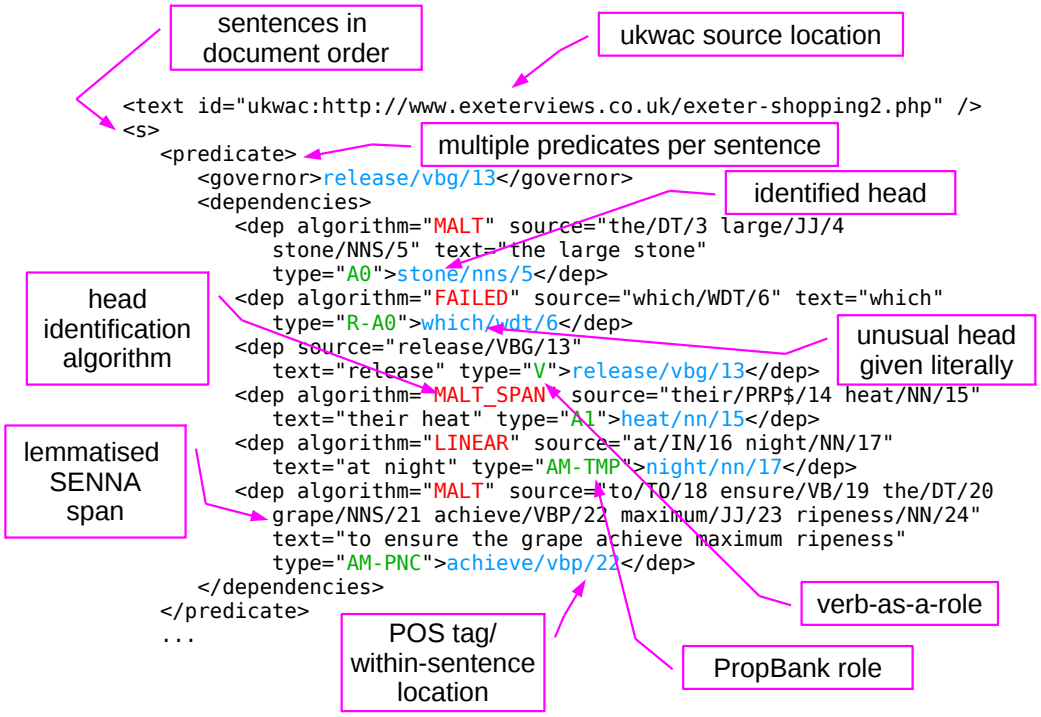

Figure 1: Excerpt of a single annotated predicate from the RW-eng corpus.

2.2. Labelling process

The labelling processing depends on the SENNA semantic role labeller. Although there are more recent and powerful labellers, we chose SENNA because it does not depend on an existing syntactic parser or heavy linguistic analysis while still maintaining reasonable accuracy (75.49% on the CoNLL 2005 task). SENNA performs labeling comparatively rapidly (36s on a Macbook i7 Pro for the CoNLL 2007 task), allowing for more flexible experimentation in corpus development over corpora the size of RW-eng. SENNA's characteristics also matched our research goals, which was to have a large-scale semantic role data source that was not completely confounded with inferences over syntactic data, as it would have been if a sophisticated parsing apparatus had been involved.

The overall procedure for labelling the corpus is as follows. For every sentence:

1. We ran the sentence through the SENNA labeller using the default model supplied with SENNA. The output consists of identified verbal predicates, and labelled spans of text connected with PropBank-style roles (Kingsbury and Palmer, 2002) to each predicate. SENNA can label overlapping stretches of text with different roles for different predicates. In our output, there is always guaranteed to be at least one verbal predicate associated with a SENNA-identified role-labelled span.

2. Each predicate verb and its associated role-labelled spans were identified in the SENNA output and grouped together.

3. Each role-labelled span was run through a heuristic process to detect the head of the span, based on a combination of part of speech tags and the MaltParser dependency trees we received with the sentences.

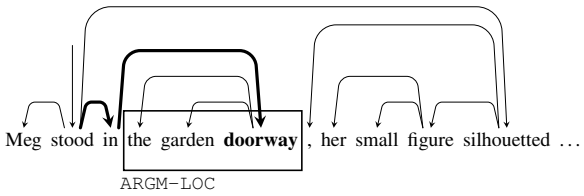

Figure 2: We illustrate the MALT heuristic with the above sentence, in which SENNA has assigned ARGM-LOC (AM-LOC in RW-eng's format) to the"the garden doorway". Transitively passing through "in", we find that "doorway" is the head, as it is the first item we encounter.

Head-detection is an issue where the nominal span is not a singleton word. This is most often the case. We take as heads only items that have nominal or verbal (open-class) POS tags. We used three heuristics for identifying the head:

- MALT – dependency links are followed from the associated verbal predicate iteratively through the dependency tree until the first word with an open-class POS tag is encountered within the bounds of the span (figure 2).

- MALT-SPAN – if MALT fails to find a verbal or nominal connection inside the span, instead look for the open-class POS-tagged word that is directly connected via the MALT parse tree to the leftmost word within the span, relative to the connections of all the other

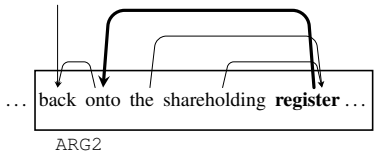

Figure 3: We illustrate the MALT-SPAN heuristic with the above ARG2-labelled phrase. The noun "register" has a connection that reaches the furthest left of all the other constituents of the span that are eligible to be identified as a head (i.e., not "onto").

neighbours. The intuition behind this is that spans for which the MALT parse does not provide a direct connection to an eligible word are often governed by a preposition or other typical left-peripheral function word (figure 3).

- LINEAR – if MALT-SPAN fails to find a qualified head, the words in the labelled span are considered iteratively from left to right using POS-based heuristics similar to those of (Magerman, 1994). More specifically, it skips over words with tags likely to be adjuncts, such as adverbs and adjectives, and when it encounters a noun or a verb, it looks ahead to make sure that it is not in an adjunct position. It returns the first noun or verb that does not have evidence, by POS tag heuristics, of being an adjunct.

The heads discovered by these heuristics are labelled as such in the output corpus, and if none of the heuristics encounter an appropriate item, it is labelled as FAILED and accepted literally as a single, full constituent.

2.3. Output format

Figure 1 contains an example sentence annotation extracted from the corpus. We describe the main features of the corpus here.

Every word in the original corpus that is mentioned in the RW-eng annotation takes the form *word/pos/N*, where *N* is the position of the word relative to the first word in the sentence, starting at 1.

Each document boundary is heralded by the text tag with the source identifier from the original corpus. Every sentence in the original corpus obtains a corresponding s tag, which contains all predicates identified by SENNA inside the sentence[3].

Each predicate tag contains one governor tag and one dependencies tag. The governor tag mentions the verb that governs the entire predicate. The dependencies tag contains a series of dep tags, which are the roles assigned by the verb and the associated text spans. A dep tag contains a source attribute, which is the tagged full span; a text attribute, which is the lemmatized text without the tagging; a type attribute, which is the Propbank-based semantic role label assigned by SENNA; and usually an algorithm attribute, which mentions one of the head-finding algorithms or FAILED if none of them worked. The verb is also repeated here as a dep tag will a type of V and no algorithm. The contents of the dep tag is the tagged word discovered by the head-finding algorithm, or the full phrase in the case of FAILED spans.

[3]Because the original corpora have access restrictions, we do not provide the original sentences, which can be obtained by matching with ukWaC and BNC. RW-eng contains word position information that only permits partial reconstruction of the original sentences; any text that SENNA does not associate with a verb requiring roles or a role-labelled text span is missing. The provided text is also lemmatized.

3. Tools

Python-based tools for manipulating the corpus exist include the aforementioned ukwac2tensor and the ukwac-heads-api[4]. The ukwac2tensor tools not only contain the scripts that convert the MALT dependency-parsed corpora to our role-labelled XML-format, but they also contain a tool to convert the XML-formatted corpus into a Pandas dataframe representing an order-3 tensor stored in HDF5 format. This tensor represents links between verbs and role-fillers found in the corpus via either counts or pointwise mutual information statistics. This representation can be used, for example, to efficiently extract role-specific feature vectors.

The ukwac-heads-api is intended for efficient access to the RW-eng data. The API contains functionality to traverse the corpus and create filtered vocabularies as well as to produce efficient Python generators for querying and sampling the corpus for applications such as deep learning. The API allows for filtering according to, for example, head-finding algorithm, and it permits iteration through randomly selected role-sets.

4. The corpus in use

RW-eng contains not only identified verbs and role-sets but also much of the underlying evidence used to detect them, such as POS tags and positional information. The head-detection as well as the full text spans of role-fillers permit the analysis of the effect of modifiers and adjuncts on thematic fit.

4.1. Corpus characteristics

RW-eng contains approximately 78 million sentences over 2.3 million documents. All together, the sentences, contain approximately 210 million identified predicates with dependent roles and 704 million identified role-fillers.

In the next section, we refer to our published results for which we used the heads discovered by our head-finding algorithm; this extrinsic evaluation of our heuristics proved to be highly successful. However, an author examined 200 of the heads in order to obtain an intrinsic estimate of how well the heuristic did, assuming the SENNA labelling as given in order to assess the heuristic alone. In 39% of the identified roles, the head-finding algorithm was not neces-

[4]https://github.com/tastyminerals/ukwac-heads-api/

Model	*Coverage* (%)	ρ
TypeDM+SDDM (Malt-only)	99	59
SDDM (Malt-only)	99	56
TypeDM	100	51
Padó	97	51
ParCos	98	48
DepDM	100	35

Table 1: SDDM and TypeDM+SDDM (Sayeed et al., 2015) are unsupervised thematic fit models based on a development version of RW-eng. These outperform other unsupervised models on a thematic fit correlation task with a common human judgement correlation task (Padó, 2007), including the very similar syntax-based TypeDM model (Baroni and Lenci, 2010).

sary because SENNA identified a single word[5]. In 48.5% of the roles, the head-finding algorithm correctly found a head in a filler with multiple words. In 4.5% of the roles, SENNA was attempting to analyze a non-sentence (such as a properly listing) and found a "gibberish" predicate and role-filler for which no head could be evaluated. 1% of the roles were multi-word expressions to which the heuristics applied a `FAILED` label. Finally, 7% of the roles were interpretable multi-word role fillers for which the heuristics identified a wrong head.

In the same sample, the `MALT` heuristic obtained a head 47% of the time. `MALT-SPAN` identified the head 11% of the time, and `LINEAR` 32%. The heuristics `FAILED` 10% of the time.

4.2. Distributional semantics and deep learning for thematic fit

We are providing RW-eng to the public domain after having developed it for our thematic fit modeling project, which has produced a number of successful results based on it. Thematic fit is the extent to which a a role-filler satisfies a given thematic role for a given predicate; it differs from selectional preferences in that it measures the extent to which a native speaker would accept the role-filler in that role, as opposed to which fillers a native speaker most expects in that role. For example, a given native speaker may expect that a secretary might be highly likely to take notes (a selectional preference), but may have a high degree of thematic fit for a doctor taking notes, even if it is not the first category of professions of which the native speaker thinks in that context. Computational models of thematic fit are typically evaluated by correlation with averaged human ratings. RW-eng has successfully been used to show that a count-based, unsupervised model of thematic fit based on PropBank roles (Sayeed et al., 2015) either outperforms a similar syntax-based model either overall or over different parts of the evaluation data; combining them produces the best-performing unsupervised models (table 1). RW-eng has also been involved in the evaluation of the role of verb senses in thematic fit modelling (Greenberg et al., 2015) and in visualization of thematic fit spaces (Sayeed et al., 2016), which is available on the web[6]. More recently, RW-eng has been used to train neural network models of thematic fit that allow for the prediction of role-fillers given varying combinations of other role-fillers, producing the best-performing models over less frequent roles like instrument and location(Tilk et al., 2016).

[5]Of these 97 items in total, 35% had ambiguous heads with possible conjoined candidates, multi-word noun compounds, ambiguous pronominal references and so on; our heuristics only return one head and thus favour precision to recall.

5. Future work

There are a number of directions in which development of this corpus could proceed, including the update and expansion of this corpus to contain more genres such as Wikipedia text. One possible future direction would be to replace SENNA with a more recent semantic role-labeller and see if the added accuracy produces better performance on our modelling tasks. However, our preliminary experiments with this have suggested that more accurate labelling produces an overfitting effect in our unsupervised models (which depend on cosine similarity in high-dimensional space).

Another area in which we are investigating the potential for improvements is in the labelling of nominal predicates (e.g., a house can fill the patient role of a sale). Our predicates are strictly based on verbs, because that is how many SRL tools including SENNA are trained. However, not only are nominal predicates frequent in language, they may be distributed differently from verbs; a resource at RW-eng's scale may be required to investigate these differences from the perspective of distributional semantics.

6. Concluding remarks

The principal effort in developing Rollenwechsel-English was that of processing and aligning heterogeneous data sources at large scale. SENNA and MaltParser produce analyses that sometimes do not directly match, and there are items in the ukWaC corpus that create practical challenges for each of these, such as special characters, accidental non-English content, specialized genres that are not sentences, and so on. However, we have derived a resource that has broad potential applications not only in psycholinguistic modeling, which was our original application, but also in other areas of semantically-aware language modeling, and we now provide it in a form that other researchers can use.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding" as well as EXC 284 Cluster of Excellence "Multimodal Computing and Interaction" during the first author's previous employment. The work reported in this paper was also supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

[6]`http://roleo.coli.uni-saarland.de/`

7. Bibliographical References

Amsel, B. D., DeLong, K. A., and Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, 82(Supplement C):118 – 132.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Collobert, R. and Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 560–567, Prague, Czech Republic, June. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of english. *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Greenberg, C., Sayeed, A., and Demberg, V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *LREC*. European Language Resources Association.

Magerman, D. M. (1994). *Natural Lagnuage Parsing as Statistical Pattern Recognition*. Ph.D. thesis, Stanford University.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7):1174–1184.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Sayeed, A., Demberg, V., and Shkadzko, P. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. In *IJCoL vol. 1, n. 1 december 2015: Emerging Topics at the First Italian Conference on Computational Linguistics*, pages 25–40. Accademia University Press.

Sayeed, A., Hong, X., and Demberg, V. (2016). Roleo: Visualising thematic fit spaces on the web. In *Proceedings of ACL-2016 System Demonstrations*, pages 139–144, Berlin, Germany. Association for Computational Linguistics.

Tilk, O., Demberg, V., Sayeed, A. B., Klakow, D., and Thater, S. (2016). Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 171–182.

Titov, I. and Klementiev, A. (2012). A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22. Association for Computational Linguistics.