Overview of AMALGUM – Large Silver Quality Annotations across English Genres

Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, Amir Zeldes
Corpling Lab
Georgetown University
{lg876, sp1184, yl879, yz565, sb1796, az364}@georgetown.edu

Introduction Corpus resources for Linguistics and NLP research on discourse phenomena, such as coreference and discourse trees, are limited by a lack of large scale, well-understood, annotated datasets: corpora are either very large (100M–10G tokens) but shallowly annotated and with unknown composition, or richly annotated, but smaller. Here, we present a resource that takes a middle path, combining some of the best features of scraped corpora – size, open licenses, lexical diversity – and high quality curated data for more interpretable inferences with complex annotations.

As a model for our resource we use the small, gold-annotated Georgetown University Multilayer (GUM) corpus (Zeldes, 2017), with 130K tokens balanced across eight genres. We scrape 4M tokens in these genres and add the same annotations available in GUM (see below), using a wide range of tools and ensembling techniques, striving for accuracy that is better than "out-of-the-box" NLP models. The resulting resource is made freely available, and is called **AMALGUM** (**A M**achine-**A**nnotated **L**ookalike of **GUM**). We envision a number of applications for the corpus, including Corpus Linguistics studies on variation, active learning, data augmentation and pretraining of NLP models.

Data We sample ∼500,000 tokens from each of 8 sources, each of which represents a genre in GUM. They are: MDPI papers (*academic*); Wikipedia biographies (*biography*); Project Gutenberg texts (*fiction*); Reddit threads (*forum*); wikiHow articles (*how-to*); Wikinews interviews (*interview*); Wikinews articles (*news*); and Wikivoyage pages (*travel*). This brings AMALGUM's size to 4M tokens, much larger than many standard benchmark corpora. Document size is 500–1K tokens, as in GUM.

Annotation & Evaluation For each document, we tokenize, tag, and lemmatize; add Universal Dependency parses and morphological features; add sentence types (declarative, imperative, question, etc.) and document structure annotations (paragraphs, headings); perform nested named and non-named entity and coreference resolution (10 entity classes, such as PERSON, PLACE etc.); and add full RST discourse parses (see the Appendix for an example). We rely on three strategies for improving annotation quality: (1) retraining tools with genre-specific data, (2) using model stacking techniques and (3) incorporating information from other layers. To assess whether these strategies help, we evaluate this approach on GUM's test set and a corrected test set of 2000 tokens from 16 AMALGUM documents balanced for genre (see Table 1).

Structural Markup All documents contain basic structural markup, including headings, paragraphs, position of figures and captions, bulleted lists, speaker information, and textual highlightings. This information is scraped directly from the source documents and is useful for subsequent layers, such as sentence splitting and discourse annotations.

Tokenization and Tagging Documents are initially tokenized using a rule-based tokenizer with postprocessing tailored to our genres. For POS tagging we train an ensemble model that takes 4 models' tag predictions as input to predict final tags. The 4 models we use here are Flair's (Akbik et al., 2019) and StanfordNLP's (Qi et al., 2018) models trained on OntoNotes (Hovy et al., 2006) and GUM. Sentence boundaries are added using an ensembled sentence splitter using XGBoost.

Dependency Parsing Universal Dependency parses and morphological features are extracted using StanfordNLP, retrained on our genres using GUM, and configured to use the tokenization and stacked POS tag predictions from the previous components, rather than rely on its own POS predictions. We use the standard StanfordNLP English model as a benchmark. The high accuracy POS

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 434-437.
Held on-line February 14-19, 2021

Tasks	Metrics	Off-the-shelf		This paper	
		AMALGUM	GUM	AMALGUM	GUM
Tokenize	F1	99.23	99.19	**99.88**	**99.89**
Tag	Accuracy	93.99	93.07	**97.37**	**97.04**
Parse	UAS / LAS	85.07 / 81.41	86.89 / 83.66	**88.81 / 85.77**	**89.47 / 85.89**
Coref	CoNLL MELA	64.4	41.4	**78.1**	**51.2**
NNER	F1	67.69	**64.41**	**71.56**	62.63
RST	S / N / R	73.93 / 46.68 / 25.06	67.62 / 43.94 / 24.17	**84.03 / 65.01 / 45.13**	**77.98 / 61.79 / 44.07**

Table 1: Comparison of performance for off the shelf tools versus our pipeline.

tagging improves the downstream parser, and our score is thus higher than the previous SOTA result on GUM test (Zeman et al., 2018).

Coreference and Nested Entity Resolution Since our gold training data is limited in size, we combine a knowledge-driven system, xrenner (Zeldes and Zhang, 2016) with contextualized BERT embeddings (Devlin et al., 2018). In preliminary experiments, we found that xrenner outperforms the SOTA coreference system on mention identification, but that many "non-coreferential" mentions (false positives) are more often incorrect for xrenner than for the SOTA nested entity recognition system (Shibuya and Hovy, 2020). We therefore created a hybrid model by injecting Shibuya and Hovy (2020)'s prediction of entity types on identical token spans, which were then fed to the modified coreference resolution system.

Discourse Parsing We add RST discourse parses to the corpus, using the DPLP parser (Ji and Eisenstein, 2014) as a benchmark. We enhanced DPLP with additional layer features: structural markup, sentence types and genre, and our high accuracy dependency trees. Discourse unit segmentation is provided by ToNy (Muller et al., 2019), and resulting trees are forced to use our sentence splits as maximal units. We also feed our system the predicted discourse function labels from a Flair sentence classifier trained on RST-DT and out-of-the-box sentiment and subjectivity scores using TextBlob's (`https://textblob.readthedocs.io/en/dev/`) pretrained model as features.

Error Analysis As the numbers from Table 1 suggest, errors are mainly found at higher levels of analysis beyong POS tagging: in parsing, entity recognition, coreference, and discourse parsing. For parsing, almost 20% of all errors come from the Reddit genre, which is likely to contain both the most vocabulary and constructions which are totally unattested in other genres. However in terms of error types, the most common problem in the Reddit data is also the most common parsing error overall: incorrect PP attachment (about 25% of errors), leading to confusion of the UD `obl` vs. `nmod` labels. The second most errorful genre, by contrast, is biographies, which has a more unique leading error type: incorrect attachment of numeric expressions, including reference numbers, dates and more, which seem to be used in ways that are untypical of other genres.

For coreference, interviews and news have the worst performance across the eight genres while fiction and voyage correctly link more than 80% of coreference relations. The genre discrepancy is likely due to the fact that existing coreference resolution systems are not good at handling coreference relations in long sentences, which are more frequent in the interviews and news. In terms of error types, the most common problem across the eights genres is paraphrasing. The system often fails to link two markables that use different words to express the same entity. In addition to that, definite expressions due to bridging anaphora and erroneous copula coreference are also typical errors in the AMALGUM test snippets.

We observed that for NER, annotation quality was similarly good across all genres, and particularly good for the academic genre. We did not observe significant systematic errors to occur: errors generally occurred only on genuinely difficult cases and had idiosyncratic causes. That said, there were some patterns that emerged. There was a bias towards categorizing entities with low-frequency words as ABSTRACT, even when this was incorrect, which we take to be reflective of ABSTRACT's disproportionately high frequency and its affinity for low-frequency words in the training data. ABSTRACT is also likely to be the most semantically diverse class, compared to the rather tightly focused PERSON or PLACE classes, meaning that while regions in embedding space for the latter types may be easier to learn, specific regions in vector space indicating the ABSTRACT class may be harder to

learn. We also found one type of error particular to fiction: mentions like "Chapter I" and "Stave II" were predicted as PERSON, most likely due to the frequent construction of a Roman numeral occurring as the suffix of person names. This shows the importance of having multiple genres in training, which could avoid the over-generalization of nominal patterns in specific genres, such as news in ACE-2005.

With regard to RST discourse parsing, the most common errors across the board are the overuse of the JOINT and ELABORATION labels, and the overuse of the PREPARATION label in fiction. Specifically, we observed the confusion of the JOINT vs. SEQUENCE labels (most often in biographies and interviews) and the ELABORATION vs. RESTATEMENT labels, and ELABORATION vs. PREPARATION labels. It is worth noting that the confusion of ELABORATION vs. PREPARATION is attributed to the incorrect parses of the nuclearity structure of the EDUs in the first place because ELABORATION is a right-branching relation label whereas PREPARATION is a left-branching one. We also found that the RST structure is heavily right-branching in fiction, voyage, and how-to guides, and with increasing depth output tends towards a stack of JOINT or SEQUENCE relations embedded into one another. Moreover, there was a small number of erroneous cases of the ATTRIBUTION label in news when no attribution verbs (e.g. *declare*, *claim*) are present, perhaps because of the high likelihood of ATTRIBUTION in news and the parser's over-reliance on the genre feature for prior label likelihood.

Overall, our error analysis suggests that cross-genre differences are a substantial challenge, which motivates the creation and utilization of genre diverse resources for NLP and quantitative linguistics research.

Discussion and Outlook Results show marked differences between off-the-shelf NLP and our tailored models with the strategies described here: retraining on in-domain data, ensembling/model stacking, and multilayer feature sharing. For the corpus presented here, this results in a substantially larger web corpus based on a small gold standard dataset with high quality results, especially for tokenization, tagging, and to a large extent, syntactic parsing. At the same time, accuracy on discourse level tasks shows that NLP tools have a long way to go before near-gold output can be expected. Realistic accuracy on new domains for tasks such as coreference resolution or NNER are well below SOTA scores, which are possible when evaluating on OntoNotes, but are less representative of what can be achieved on web data "in the wild". In future work we plan to enhance corpus quality, using active learning, bootstrapping, and targeted use of crowdsourcing, with the aim of especially improving discourse level annotations, such as coreference and discourse parsing.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of NAACL 2019*, pages 724–728.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL 2006*, pages 57–60, New York.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL 2014*, pages 13–24, Baltimore, MD.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of DISRPT 2019*, pages 115–124, Minneapolis, MN.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of CoNLL 2018*, pages 160–170, Brussels, Belgium.

Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *TACL*, 8:605–620.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of CORBON 2016*, pages 92–101.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task. In *Proceedings of CoNLL 2018*, pages 1–21, Brussels, Belgium.

A Appendix: Sample Analyses

The figures below visualize predicted output for entity recognition, coreference resolution and discourse parsing for one of the samples from the AMALGUM test set.

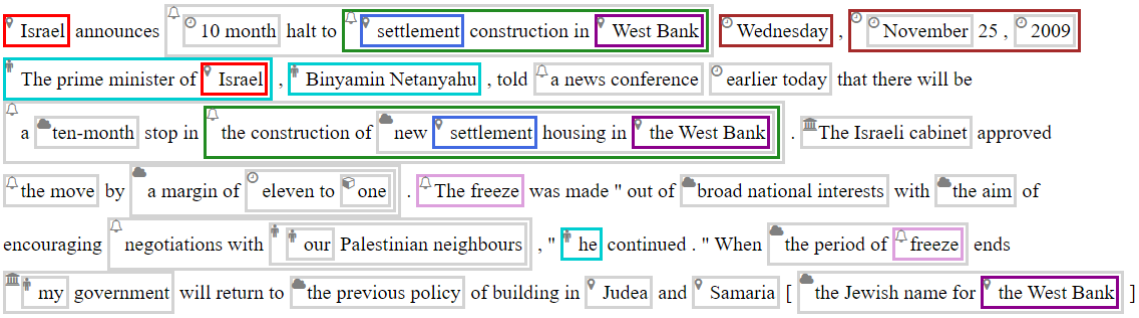

Figure 1: Xrenner's coreference and entity predictions on an AMALGUM news snippet. Coreferent mentions are colored (e.g. Israel and the prime minister of Israel are boxed in red and cyan), and entity types are indicated by icons: PLACE, PERSON, TIME, EVENT, ABSTRACT, and OBJECT.

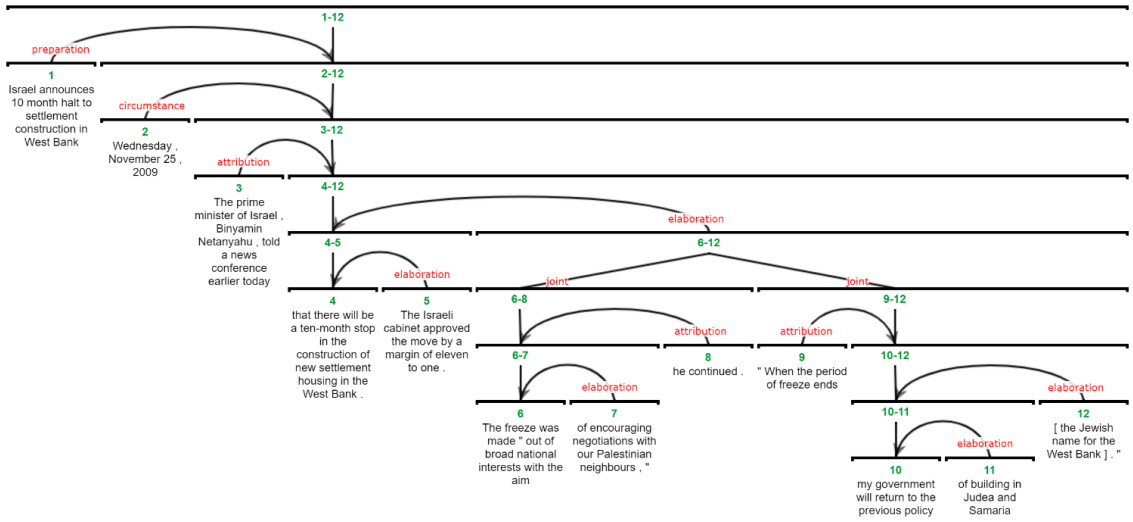

Figure 2: Predicted discourse parse for the same news text; errors include viewing the circumstance clause 'when the freeze ends' as ATTRIBUTION and an incorrect attachment of the ELABORATION about the name of the West Bank. In contrast to a human analyst's manual annotations, the parser also groups units [6-12] as a JOINT, which is not implausible.