

Text-based NP Enrichment

Yanai Elazar* Victoria Basmov* Yoav Goldberg Reut Tsarfaty

Computer Science Department, Bar Ilan University

Allen Institute for Artificial Intelligence, Israel

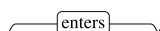
{yanaiela, vikasaeta, yoav.goldberg, reut.tsarfaty}@gmail.com

Abstract

Understanding the relations between entities denoted by NPs in a text is a critical part of human-like natural language understanding. However, only a fraction of such relations is covered by standard NLP tasks and benchmarks nowadays. In this work, we propose a novel task termed *text-based NP enrichment* (TNE), in which we aim to enrich each NP in a text with all the preposition-mediated relations—either explicit or implicit—that hold between it and other NPs in the text. The relations are represented as triplets, each denoted by two NPs related via a preposition. Humans recover such relations seamlessly, while current state-of-the-art models struggle with them due to the implicit nature of the problem. We build the first large-scale dataset for the problem, provide the formal framing and scope of annotation, analyze the data, and report the results of fine-tuned language models on the task, demonstrating the challenge it poses to current technology. A webpage with a data-exploration UI, a demo, and links to the code, models, and leaderboard, to foster further research into this challenging problem can be found at: yanaiela.github.io/TNE/.

1 Introduction

A critical part of understanding a text is detecting the entities in the text, denoted by NPs, and determining the different semantic relations that hold between them. Some semantic relations between NPs are explicitly mediated via verbs, as in (1):

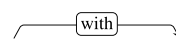
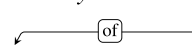
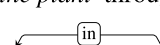
- (1)  Water enters the plant through the roots.

Much work in NLP addresses the recovery of such verb-mediated relations (SRL) (Gildea and Jurafsky, 2002; Palmer et al., 2010), either using pre-specified role ontologies such as PropBank or FrameNet (Palmer et al., 2005; Ruppenhofer et al., 2016), or, more recently, using natural-language-based representations (QA-SRL) (He et al., 2015;

*Equal contribution.

FitzGerald et al., 2018). Another well-studied kind of semantic relations between NPs is that of *coreference* (Vilain et al., 1995; Pradhan et al., 2012), where two (or more) NPs refer to the same entity.

Such NP-NP relations, which are either mediated by verbs (as in SRL or Relation Extraction) or form coreference relations, represent only a subset of the NP-NP relations that are naturally expressed in texts. Consider, for instance, the following sentences:

- (2)  A person with brown eyes crossed the street.
(3)  Water enters the plant through the roots.
(4)  I entered the room, the window was open.

All of the above cases contain examples of NP-NP relations, where the type of relation can be expressed via an English preposition. The preposition may be explicit in the text, as in (2), where the relation **A person** with **blue eyes** is explicitly expressed, or they may be *implicit* and left to the reader to infer, as in (3)-(4); in (3) readers easily infer that **the roots** are of **the plant**. Likewise, in (4) readers infer that **the window** is in **the room**.¹ Properly understanding the text means knowing that these relations hold, even when they are not explicitly stated in the utterance. Figure 1 shows additional examples.

These relations, both explicit and implicit, are indispensable for understanding the text. While human-readers infer these relations intuitively and spontaneously while reading, machine-readers generally ignore them. In this work, we thus propose a new NLU task in which we aim to recover all the preposition-mediated relations—whether explicit or implicit—between NPs that exist in a text. We call this task *Text-based NP Enrichment* or *TNE* for short.

¹Here, both “in” and “of” are possible prepositions, but “in” is slightly more specific.

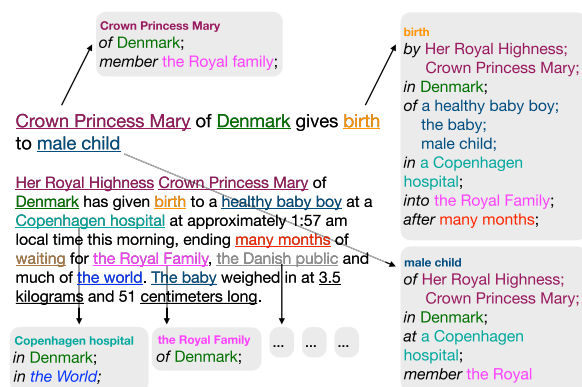


Figure 1: Preposition-mediated relations between NPs in a text. NPs in the same color designate the same entity (co-refer). Gray boxes show all the preposition-mediated relations for a single NP anchor (some are indicated with “...” for brevity). This figure shows a title and a single short paragraph. The texts in our dataset span 3 paragraphs.

The short examples (2)–(4) that illustrate the phenomenon may not look challenging to infer using current NLP technology. However, when we go beyond sentence level to document level, things become substantially more complicated. As we demonstrate in §6, a typical 3-paragraph text in our dataset has an average of 35.8 NPs, which participate in an average of 186.7 preposition-mediated relations, the majority of which are implicit. Figure 2 shows a complete annotated document from our dataset.

The type of information recovered by the NP Enrichment task complements well-established core NLP tasks such as entity typing, entity linking, coreference resolution, and semantic-role labeling (Jurafsky and Martin, 2009). We believe it serves as an important and much-needed building block for downstream applications that require text understanding, including information retrieval, relation extraction and event extraction, question answering, and so on. In particular, the NP Enrichment task *neatly encapsulates much of the long-range information* that is often required by such applications. Take for example a system that attempts to extract reports on police shooting incidents (Keith et al., 2017), with the following challenging, but not uncommon, passage:²

Police officers spotted the butt of a handgun in Alton Sterling’s front pocket

²We thank Katherine Keith for this example.

and saw him reach for the weapon before opening fire, according to a Baton Rouge Police Department search warrant filed Monday that offers the first police account of the events leading up to his fatal shooting.

Considering this shooting-event passage, an ideal coreference model will resolve *his* to *Alton Sterling’s*, making the entity being shot local to the shooting event. On top of that, an ideal NP Enrichment model as we propose here will also recover:

fatal shooting [of Alton Sterling] [by Police officers [of Baton Rouge Police Department]]

making the shooter identity local to the shooting event as well, ready for use by a downstream event-argument extractor or machine reader.

Of course, one could hope that a dedicated, end-to-end-trained shooting-events extraction model will learn to recover such information on its own. However, it will require pre-defining the frame of *shooting* events, and it will require a substantial amount of training data to get it right (which often does not happen in practice). Focusing on *Text-based NP Enrichment* provides an opportunity to learn a core NLU skill that does not focus on a pre-defined set of relations, and is not specific to a particular benchmark. Finally, beyond its potential usefulness for downstream NLP applications, the Text-based NP Enrichment task serves as a *challenging benchmark for reading comprehension*, as we further elaborate in §3.

In what follows we formally define the *Text-based NP Enrichment* task (§2) and its relation to *reading comprehension* (§3), we describe a large-scale high-quality English TNE dataset we collected (§4) and its curation procedure (§5). We analyze the dataset (§6) and experiment with pre-trained language model baselines (§7), achieving moderate but far-from-perfect success on this dataset (§8). We also conduct an analysis of the best model, showcasing the strengths, weaknesses, and open challenges of the best model (§9). We then discuss the relation of TNE to other linguistic concepts, such as *bridging* (Clark, 1975), *relational nouns* (Partee, 1983/1997; Loebner, 1985; Barker, 1995), and *implicit arguments* (Ruppenhofer et al., 2009; Meyers et al., 2004; Gerber and Chai, 2012; Cheng and Erk, 2019) (§10). We finally

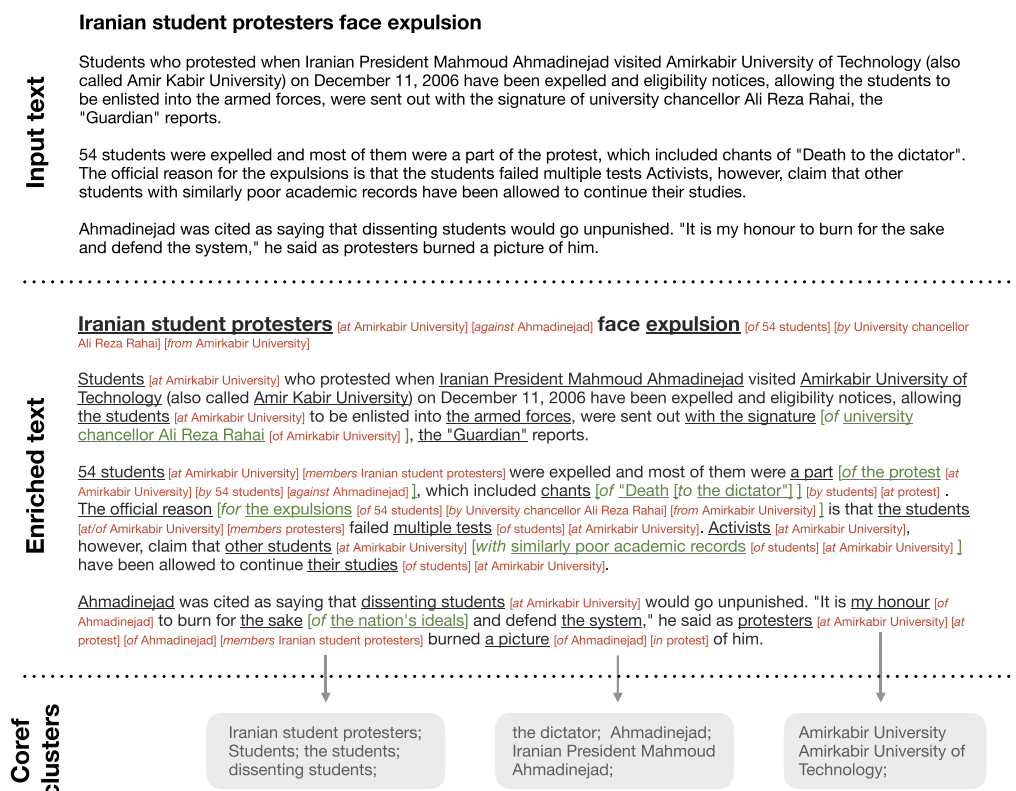


Figure 2: NP-enriched document from the dataset. The title appears in a larger font, the NPs in the document are marked with underline. The **green** NP-enrichments appear explicitly in the original text, while the **red** do not, and are typically harder to infer. For brevity, each link in the text mentions only one of the NPs in a coreference cluster. The dataset has additional links to the other NPs in the cluster.

conclude that, in contrast to those linguistic tasks, the *Text-based NP Enrichment* task is more exhaustive, sharply scoped, easier to communicate, and substantially easier to consistently annotate and use by non-experts.

2 Text-based NP Enrichment (TNE)

Task Definition The Text-based NP Enrichment task is deceptively simple: For each ordered pair (n_1, n_2) of non-pronominal base-NP³ spans in an input text, determine if there exists a preposition-mediated relation between n_1 and n_2 , and if there is one, determine the preposition that best describes their relation.⁴ The output is a list

³We follow the definition of Base-NPs as defined by Ramshaw and Marcus (1995): initial portions of non-recursive noun-phrases, including pre-modifiers such as determiners, adjectives and noun-compounds, but not including post-modifiers such as prepositional phrases and clauses. These are also known in the NLP literature as “NP Chunks”.

⁴During annotation, we noticed that annotators often tried to express set-membership using prepositions, which resulted in awkward and unclear annotations. To remedy this, we found it effective to add an explicit “member-of” relation as an allowed annotation option. This significantly reduced

of tuples of the form (n_i, prep, n_j) , where n_i is called the *anchor* and n_j is called the *complement* of the relation. Figure 2 shows an example of text where each NP n_1 is annotated with its (prep, n_2) NP-enrichments.

Despite the task’s apparent simplicity, the underlying linguistic phenomena are quite complex, and range from simple syntactic relations to relations that require pragmatics, world knowledge, and common-sense reasoning. Performing well on the task suggests a human-like level of understanding. Notably, human readers detect most of the relations almost subconsciously when reading, while some of the relations require an extra conscious inference step.

Example Consider the following text:

(5) *Adam’s father went to meet the teacher at his school.*

In this utterance, there are four non-pronominal base-NPs: “Adam”, “father”, “the teacher” and “his school”. The task requires identifying the relations between these NPs, such as “father” being the parent of “Adam”, “the teacher” being the person “Adam’s father” went to meet, and “his school” being the school of “the teacher”. This task is designed to improve the consistency and quality of the annotation. While not officially part of the task, we do keep these annotations in the final dataset.

“his school”, which makes 12 possible pairs: (**Adam**, *father*), (**Adam**, *the teacher*), . . . , (**his school**, *the teacher*).

The preposition-mediated relations to be recovered in this example are:⁵

- i (**father**, *of Adam*)
- ii (**the teacher**, *of Adam*)
- iii (**the teacher**, *at his school*)
- iv (**his school**, *of Adam*)

The first items are anchors, and the latter ones are the complements.

Order The order of appearance of NPs within the text does not matter: For a given pair of NPs n_1 and n_2 , we consider both (n_1, n_2) and (n_2, n_1) as potential relation candidates, and it is possible that both relations will hold (likely with different prepositions). The only restriction is that an NP span cannot relate to itself. For a text with k NPs, this results in $k^2 - k$ candidate pairs.

Scope In terms of the annotated relations, we are interested in the set of semantic relations that can be expressed in natural language via the use of a preposition. This identifies a rich, cohesive and well-scoped set of NP-NP relations that are not mediated by a verb and are not coreference relations. Importantly, we restrict ourselves to NPs that are mentioned in the text, excluding relations with NPs that reside in some text-external shared context. For example, consider the sentence: “*The president discussed the demonstrations near the border*”. Here, the NPs “*the president*”, “*the border*”, and “*the demonstrations*” are all under-determined, and, to be complete, should relate to other NPs using preposition-mediated relations: **president** [*of Country*]; **border** [*of CountryX*] [*with CountryY*]; **demonstration** [*by some-group*] [*about some-topic*]. However, as these complement NPs do not appear in the text, we do not consider them to be part of the TNE task.

⁵We note that some of these are ambiguous. For example, it is not 100% certain that the teacher is indeed “the Teacher of Adam”. For example, it could be a teacher of Adam’s sister. Yet, without further information, many if not most readers will interpret it as the teacher of Adam. This kind of ambiguity is an inherent property of language, and—like in many other datasets, cf. NLI—we deliberately opted for wide-coverage over preciseness: We are interested in what a “typical human” might infer as a relation, and not only 100% certain ones.

The Use of Prepositions as Semantic Labels

While the relations we identify between NPs can be expressed using prepositions, one could argue that using prepositions as semantic labels is not ideal, due to their inherent ambiguity (Schneider et al., 2015, 2016, 2018; Gessler et al., 2021): indeed a preposition such as *for* has multiple senses, and can indicate a large set of semantic relations ranging from BENEFICIARY to DURATION.

We chose to use prepositions as relation labels, despite this ambiguity. This follows a line of annotation work that aims to express semantic relations using natural language (FitzGerald et al., 2018; Roit et al., 2020; Klein et al., 2020; Pyatkin et al., 2020), as opposed to works that used formal linguistic terms, traditionally relying on expert-defined taxonomies of semantic roles and discourse relations. The aforementioned works label predicate-argument relations using restricted questions. In the same vein, we label nominal relations using prepositions.

We argue that the preposition-based labels are useful for humans and machines alike: Humans can easily understand the task (both as annotators and—perhaps more importantly—as consumers), and current machine learning models are quite effective with implicitly dealing with prepositions ambiguity.⁶ Moreover, while the prepositions themselves are ambiguous, the (NP, prep, NP) triplet provides context that is, in many cases, sufficient to disambiguate the coarse-grained preposition sense.

We find that the preposition-based annotation has the following advantages: It clearly scopes the task with respect to the kinds of relations that are contained in it; and it is expressive, capturing a large class of interesting semantic relations. On top of that, the task and the corresponding relation-set is easy to explain to both human annotators (thus allowing to obtain high levels of agreement) and to human consumers of the model (allowing wider adoption, as the task and its output does not require special training to understand). Finally,

⁶For example, in the SQUAD question-answering dataset (Rajpurkar et al., 2016), 15% of all questions either begin or end with a preposition, and 30% of all answer spans directly follow a preposition, requiring the models to deal with their ambiguous nature in order to perform well. The high accuracy scores obtained on the SQUAD dataset indicates that the models indeed succeed in the face of ambiguity. Relation-extraction tasks also work to a large extent around prepositional phrases, and manage to effectively extract relations.

the output can be easily fed into existing NLP systems, which already deal to a large extent with the inherent ambiguities of prepositions and prepositional phrases.

To conclude, we argue that despite the ambiguities of prepositions, they allow us to obtain a meaningful set of *typed* semantic links between NPs, which are well understood by people and can be effectively processed by NLP models. While the annotation can be refined to include a fine-grained sense annotation for each link, for example, via a scheme as that of Schneider et al. (2018), we leave such an extension to future work.

Coreference Clusters A common relation between NPs is that of *identity*, also known as a *coreference* relation, where two or more NPs refer to the same entity. How do coreference relations relate to the NP Enrichment task? While the NP Enrichment task so far is posed as inferring prepositional relations between NPs, in actuality the prepositional relations hold between an NP and a coreference cluster. Indeed, if there is a prepositional relation $\text{prep}(n_1, n_2)$, and a coreference relation $\text{coref-to}(n_2, n_3)$, we can immediately infer the link $\text{prep}(n_1, n_3)$.⁷ We make use of this fact in our annotation procedure, and the dataset includes also the coreference information between all NPs in the text. Indeed, for brevity, Figure 2 shows only a subset of the relations, indicating for each anchor NP only a single complement NP from each coreference cluster. Some of the coreference clusters are shown at the bottom of the Figure. Note that the coreference clusters are not part of the task’s input or expected output.

Formal Dataset Description An input text is composed of tokens w_1, \dots, w_t , and an ordered set $N = n_1, \dots, n_k$ of base-NP mentions. The underlying text is often arranged into paragraphs, and may also include a title. A base-NP mention,

⁷Note that the converse does not hold: $\text{prep}(n_1, n_2)$, $\text{coref-to}(n_1, n_3)$ does not necessarily entail $\text{prep}(n_3, n_2)$. Consider for example: “*The race began. John, the organizer, pleased*”. While *John* and *the organizer* are coreferring, the relation **organizer** *of the race* holds, while **John** *of the race* does not. This is because *John* and *the organizer* are two different senses for the same reference, and the relation holds only for one of the senses (cf. Frege, 1960). Putting it differently, when *John* and *organizer* serve as predicates, their selectional preferences are different despite them coreferring. Such examples are common, consider also “*John is Jenny’s father, Mary’s husband*” where **father** *of Jenny* holds, while **husband** *of Jenny* doesn’t. Similarly, **husband** *of Mary* holds, while **father** *of Mary* doesn’t.

also known as NP chunk, is the smallest noun phrase unit that does not contain other NPs, prepositional phrases, or relative clauses.⁸ It is defined as a contiguous span over the text, indicated by start-token and end-token positions (e.g., (3, 5) “the young boy”). The output is a set R of relations of the form (n_i, prep, n_j) , where $i \neq j$ and prep is a preposition (or a set-membership symbol). Each text is also associated with a set C of non-overlapping coreference clusters, where each cluster $c \subseteq N$ is a non-empty list of NP mentions. The set of clusters is *not* provided as input, but for correct sets R it holds that $\forall n_{j'} \in c(n_j), (n_i, \text{prep}, n_j) \in R \Rightarrow (n_i, \text{prep}, n_{j'}) \in R$, where $c(n_j) \in C$ is the cluster containing n_j .

Completeness and Uniformity The kinds of preposition-mediated relations we cover originate from different linguistic or cognitive phenomena, and some of them can be resolved by employing different linguistic constructs. For example, some within-sentence relations can be extracted deterministically from dependency trees, for example, by following syntactic prepositional attachment. Other relations can be inferred based on pronominal coreference (e.g., “his school [of Adam]” above can be resolved by first resolving “his” to “Adam’s” via a coreference engine, and then normalizing “Adam’s school” \rightarrow “school of Adam”). Many others are substantially more involved. We deliberately chose not to distinguish between the different cases, and expose all the relations to the user (and to the annotators) via the same uniform interface. This approach also contributes to the practical usefulness of the task: Instead of running several different processes to recover different kinds of links, the end-user will have to run only one process to obtain them all.

Evaluation Metrics Our main metrics for evaluating NP enrichment tasks are precision, recall, and F1 on the recovered triplets (links) in the document. For analysis, we also report two additional metrics: precision/recall/F1 on unlabeled links (where the preposition identity does not matter), and accuracy of predicting the right preposition when a gold link is provided. We break this last metric into two quantities: accuracy of predicting the preposition for gold links that were recovered by the model, and accuracy of prepositions for gold links that were not recovered.

⁸We use an automatic parser to obtain such base-NPs, using spaCy’s parser (Honnibal et al., 2020).

3 TNE as a Reading Comprehension Benchmark

While *reading comprehension* (RC) and *question answering* (QA) are often used interchangeably in the literature, measuring the reading comprehension capacity of models via question answering, as implemented in benchmarks such as SQuAD (Rajpurkar et al., 2016), BoolQ (Clark et al., 2019) and others, has several well-documented problems (Dunietz et al., 2020). We argue that the TNE task we propose herein has properties that make it appealing for assessing RC, more than QA is.

First, benchmarks for extractive (span-marking) QA are sensitive to the span-boundary selection, on the other hand, benchmarks for yes/no, multiple choice, or generative questions can in principle be answered in a way that is completely divorced from the text. On a more fundamental level, all QA benchmarks are very sensitive to lexical choices in the question and its similarity to the text. Furthermore, QA benchmarks rely on human authored questions that are easy to solve based on surface artifacts. Finally, in many cases, the existence of the question itself provides a huge hint towards the answer (Kaushik and Lipton, 2018).

The underlying cause for all of these issues is that QA-based setups do not measure the comprehension of a *text*, but rather comprehending a (*text*, *question*) pair, where the question adds a significant amount of information, focuses the model on specific aspects of the text, and exposes the evaluation to biases and artifacts. The reliance on the human-authored questions makes QA a bad format for measuring “text understanding”—we are likely measuring something else, such as the ability of the model to discern patterns in human question-writing behavior.

The TNE task we define side-steps all the above issues. It is based on the text alone, without revealing additional information not present in the text. The exhaustive nature of the task entails looking both at positive instances (where a relation exists) and negative ones (where it doesn’t), making it harder for models to pick up shallow heuristics. We don’t reveal information to a model, beyond the information that the two NPs appear in the same text. Finally, the list of NPs to be considered is pre-specified, isolating the problem of *understanding the relations* between NPs in the text from the much easier yet intervening problem of *identifying NPs* and *agreeing on their exact spans*.

of, against, in, by, on, about, with, after, to, from, for, among, under, at, between, during, near, over, before, inside, outside, into, around

Table 1: Prepositions used in TNE.

Thus, we consider TNE a less biased and less gameable measure of RC than QA-based benchmarks. Of course, the information captured by TNE is limited and does not cover all levels of text understanding. Yet, performing the task correctly entails a non-trivial comprehension of texts, which human readers do as a byproduct of reading.

4 Text-based NP Enrichment Dataset

We collect a large-scale TNE dataset, consisting of 5.5K documents in English (3,988 train, 500 dev, 500 in-domain test, and 509 out-of-domain test). It covers about 200K NPs and over 1 million NP relations. The main domain is WikiNews articles, and the out-of-domain (OOD) texts are split evenly between reviews from IMDB, fiction from Project Gutenberg, and discussions from Reddit.

Each annotated document consists of a title and 3 paragraphs of text, and contains a list of non-pronominal base-NPs (most identified by SpaCy [Honnibal et al., 2020]⁹ but some added manually by the annotators), a list of coreference clusters over the NPs, and a list of NP-relations that hold in the text. Each relation is a triplet consisting of two NPs from the NP list, and a connecting element which is one of 23 prepositions (displayed in Table 1)¹⁰ or a “member(s) of” relation designating set-membership. The list of NP relations is *exhaustive*, and aims to cover all and only valid NP-NP relations in the document.

5 Data Annotation and Curation

5.1 Annotation Procedure

We propose a manual annotation procedure for collecting a large-scale dataset for the TNE task. Considering all $k^2 - k$ NP pairs (with an average k of 35.8 in our dataset) is tedious, and, in our experience, results in mistakes and inconsistencies. In order to reduce the size of the space and improve annotation speed, quality, and consistency,

⁹v.3.0.5, model *en_core_web_sm*.

¹⁰The set was initiated with the 20 most common prepositions in English, and we added three additional prepositions that were requested during the initial annotation phase.

we opted for a **two-stage process**, where the first stage includes the annotation of coreference clusters over mentions, and the second stage involves NP Enrichment annotation over the clusters from the first stage. We find that this two-stage process dramatically reduces the number of decisions that need to be taken, and also improves recall and consistency by reducing the cognitive load of the annotators, focusing them on a specific mode at each stage. We hereby describe the different stages.

Stage 1: Annotating Coreference Clusters

We start by collecting coreference clusters, as well as discarding non-referring NPs, that are “irrelevant” for the next stage (such as time-expressions). We created a dedicated user-interface to facilitate this procedure (Figure 3a). The annotators go over the NPs in the text in order, and, for each NP, indicate if it is (a) a new mention (forming a new cluster); (b) “same as” (corefering to an entity in an existing cluster initiated earlier); (c) a time or measurement expression; or (d) an idiomatic expression. At each point, the annotators can click on a previous NP to return to it and revise their decisions.

The OOD and documents from the test-set were annotated by two annotators for measuring agreement. They were then consolidated by one of the paper’s authors for high-quality annotations.

Stage 2: Annotating NP-relations The second step is the NP Enrichment relation annotation. The annotators are exposed to a similar interface (Figure 3b). For each NP, they are presented with all the coreference clusters, and must indicate for each cluster if there is a preposition-mediated relation between the NP and the cluster.

For this stage, all documents are annotated by two annotators and undergo a consolidation step. The **consolidation** over the two annotators is performed by a third annotator, who did not see the document before. This annotator is presented with the interface shown in Figure 3c. The consolidator sees all the relations created by the two preceding annotators, and decides which of them are correct.^{11,12}

¹¹We measure the agreement of this step by an additional annotators that consolidate 10% of the documents, and report the agreement between the two consolidators.

¹²In this stage, two links with identical NPs can be chosen with different prepositions (e.g., Ex. (4)). This may increase the number of possible relations in a given document from $k^2 - k$ possible pairs to $(k^2 - k) * p$, where p is the number of considered prepositions. However, in practice, having more

(a) Coreference clusters collection interface.

(b) NP Enrichment data collection interface.

(c) NP Enrichment consolidation interface.

Figure 3: Interfaces of the annotation steps.

5.2 Annotators

We trained and qualified 23 workers on the Amazon Mechanical Turk (AMT) platform, to participate in the coreference, NP relations, and consolidation tasks. We follow the controlled crowdsourcing protocol suggested by Roit et al. (2020) and Pyatkin et al. (2020), giving detailed

than two prepositions for the same NP pairs is not common, and two prepositions occur in 11.6% of the test-set. For simplicity, in this work, we consider a single preposition for each NP pair, but the collected data may contain two prepositions for some pairs.

	In-Domain		Out-of-Domain				all
	train	test	Books	IMDB	Reddit	OOD	
CoNLL (Coref)	–	82.1	76.8	77.6	78.6	77.1	79.8
Relation-F1	89.8	94.4	87.0	89.6	90.2	88.9	90.3
IPrep-Acc	99.8	100.0	99.5	99.8	100.0	99.8	99.9
UPrep-Acc	100.0	100.0	100.0	100.0	100.0	100.0	100.0
F1	89.6	94.4	86.6	89.4	90.2	88.6	90.1

Table 2: Agreement scores on the different annotation parts. We report both the coreference CoNLL scores, and the metrics of NP Enrichment calculated on the consolidated annotations.

instructions, training the workers, and providing them with ongoing personalized feedback for each task.

We paid \$1.50, \$2.50, and \$1.5) for each HIT in the coreference, NP-relations, and consolidation tasks, respectively. The price for the NP-relations task was raised to \$2.70 for the test and out-of-domain subsets. We additionally paid bonus payments on multiple occasions. Overall, we aimed at paying at least the minimum wage in the United States.

5.3 Inter-annotator Agreement

We report the agreement scores for the coreference and the consolidated relation annotations. The full results, broken by split, are reported in Table 2. The IPrep-Acc and UPrep-Acc metrics measure the preposition-only agreement (whether the annotators chose the same preposition for a given identified NP-pair), and are discussed in §9.1.

Coreference We follow Cattan et al. (2021) and evaluate the coreference agreement scores after filtering singleton clusters. We report the standard CoNLL-2012 score (Pradhan et al., 2012) that combines three coreference metric scores. The in-domain test score¹³ is 82.1, while in the OOD the score is 77.1. For comparison with the most dominant coreference dataset, OntoNotes (Weischedel et al., 2013), which only reported the MUC agreement score (Grishman and Sundheim, 1996), we also measure the MUC score on our dataset. The MUC score on our dataset is 83.6, compared to 78.4-89.4 in OntoNotes, depending on the domain (Pradhan et al., 2012). It is worth

¹³To reduce costs and time, we did not collect double annotation for the train split, thus we cannot report agreement on it.

noting that on the Newswire domain of OntoNotes (Weischedel et al., 2013) (the domain that is most similar to ours) the score is 80.9, which indicates a high quality of annotation in our corpus. We expect the quality of our final coreference data to be even higher due to the consolidation step that was done by an expert on the test set and OOD splits.

NP-relations Next, we report agreement scores on the NP-relations consolidation annotation, which were measured on 10% of all the annotations. We use the same metrics for the NP Enrichment task (§2) and use one of the annotations as gold, and the other as the prediction. Thus we only report accuracy and F1 scores (the precision and recall are symmetric depending on the role of each document). The Relation-F1 scores for the train and test are 89.8 and 94.4 respectively, while for the OOD it is 88.9. The preposition scores are almost perfect in all splits, with an average of 99.9 when the annotators agree on the link and 100.0 when they don’t. Finally, the F1 scores also differ between splits: 89.6, 94.4, and 88.6 for the train, test, and OOD, respectively, but are overall high.

6 Dataset Statistics and Analysis

We report statistics of the resulting NP Enrichment dataset, and summarize them in Table 3. Overall, we collected 5,497 documents, with per-document averages of 35.8 NPs, 5.2 non-singleton coreference clusters, and 186.7 NP-relations. The average number of tokens in a document is 163.3 tokens, where the largest document has 304 tokens.

Distribution of Prepositions We analyze the prepositions in the relations we collected. We aggregate the prepositions of the test set from all relations and present their distribution in Figure 4. We only show prepositions that appear at least in 4% of the data, and the rest are aggregated together into the *Other* label. The most common preposition is *of*, followed by *in*, which constitute 23.9% and 19.8% of the prepositions in our data, respectively. The rest of the prepositions are used much less frequently, with *from* and *for* appearing in 9.7% and 6.3% of the prepositions, respectively. The least used preposition is *into*, which appears in 0.07% of the prepositions.

NP-relations We provide some statistics that shed light on the nature of the preposition-mediated NP-NP relations in the annotated data.

	In-Domain			Out-of-Domain				all
	train	dev	test	Books	IMDB	Reddit	OOD-all	
Documents	3,988.0	500.0	500.0	170.0	169.0	170.0	509.0	5,497.0
Tokens	651,835.0	81,741.0	77,618.0	30,133.0	29,285.0	27,181.0	86,599.0	897,793.0
NPs	143,406.0	17,815.0	17,521.0	6,502.0	6,099.0	5,803.0	18,404.0	197,146.0
NP-Links	744,513.0	103,668.0	120,198.0	22,886.0	25,164.0	10,228.0	58,278.0	1,026,657.0
Coref-Clusters	21,473.0	2,598.0	2,581.0	773.0	759.0	821.0	2,353.0	29,005.0
Coref-Links	354,734.0	51,776.0	56,443.0	11,847.0	11,798.0	5,347.0	28,992.0	491,945.0
Avg. Surface Distance	53.9	52.8	52.6	53.6	58.2	47.9	54.6	53.7
Avg. Symmetric Links	10.2	12.6	14.3	6.7	26.5	1.7	11.6	10.9
Avg. Transitive	95.5	119.2	144.2	52.8	64.5	14.8	44.0	97.3
% Title Links	13.6	12.7	12.0	8.3	13.7	25.4	15.8	13.6
% Backward-relations	56.9	56.0	55.7	56.7	56.6	57.3	56.8	56.7
% Surface-Form	4.0	3.5	2.9	4.3	2.7	5.4	4.1	3.9
% Surface-Form+	6.3	5.6	4.6	6.8	3.5	7.2	5.9	6.0
% Intersentential	84.4	85.1	85.5	79.8	87.7	83.2	83.8	84.6

Table 3: Statistics summary of the NP Enrichment dataset.

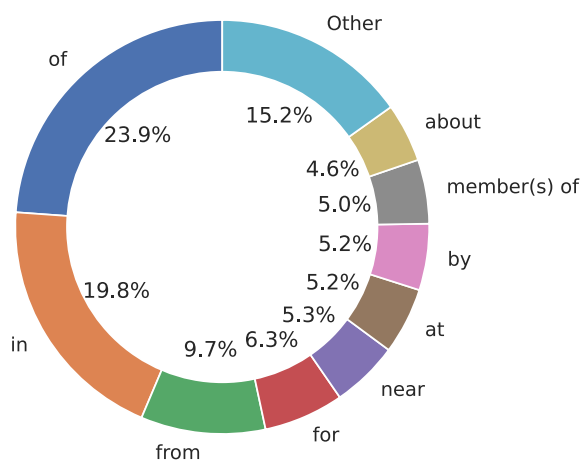


Figure 4: Distribution of the prepositions in the NP Enrichment test set.

First, we measure the *surface distance* between NPs in the relations, in terms of token counts between the *anchor* and the *complement*. We found the average distance to be 53.7 tokens, indicating an average large distance between two NPs, which demonstrates the task’s difficulty. *Backward-relations* (as opposed to *forward-relations*) are relations where the *complement* appears before the *anchor*—56.7% of the relations are backward. Sometimes, the string “**anchor** preposition *complement*” appears directly in the text. We call these cases *Surface-Form*. For instance in Ex (5) (**the teacher** at *his school*) is a *Surface-Form* relation. We computed the percentage of such relations in the data and found only 3.9% of them to be of such type. We also relax this definition and search for the preposition following the complement in a window size of 10 from the anchor, which we

call *Surface-Form+*. The percentage of such cases remains low: 6.0% of the links. *Symmetric* relations are two relations between the same two NPs, that differ in direction (and potentially the preposition). For instance, in Figure 8, the following links are symmetric (**website**, of, *the owners*) and (**the owners**, of, *website*). On average, there are 10.9 such symmetric relations in a document. Finally, *transitive* relations are sets of three NPs, *a*, *b*, and *c*, that include relations between (*a*, *b*), (*b*, *c*), and (*a*, *c*) (the preposition identity is not relevant). We found an average of 97.3 transitive relations per document in total.

Explicit vs. Implicit NP Relations Next, we analyze the composition of the relations in the dataset, as to whether these relations are implicit or explicit. While there is no accepted definition of explicit-implicit distinction in the literature (Carston, 2009; Jarrah, 2016), here we adapt a definition originally used by Cheng and Erk (2019) for another phenomenon, implicit arguments:¹⁴ In an *implicit* relation the *anchor* and the *complement* are not syntactically connected to each other and might not even appear in the same sentence. This implies, for example, that any inter-sentential relations are implicit¹⁵, while relations within one sentence can be either implicit or explicit. We sample three documents from the test-set, containing 590 links in total, and count the number of relations of each type. Our manual analysis reveals that 89.8% of the relations are implicit.

¹⁴Implicit arguments “are not syntactically connected to the predicate and may not even be in the same sentence” (Cheng and Erk, 2019).

¹⁵84.6% of all the links in our dataset are inter-sentential.

Bridging vs. TNE *Bridging* has been extensively studied in the past decades, as we discuss in §10. Here, we explore how many of the relations we collected correspond to the definition of *bridging*. We use the same three documents from the analysis described above, and follow the annotation scheme from ISNotes1.0 (Markert et al., 2012)¹⁶ to annotate them for *bridging*. We found that 15 out of the 590 links (2.5%) in these documents are *bridging* links (i.e., meet the criteria for bridging defined in ISNotes). These three documents contain 104 NPs, that is, the ratio of bridging links per NP is 0.14. While the ratio is small, it is larger than the ratio in ISNotes, which contains 663 bridging links out of 11K annotated NPs (Hou et al., 2013b), that is, 0.06 bridging links per NP.

7 Deterministic Baselines

We explore multiple deterministic baselines, which should expose regularities in the data that models may use (and therefore may result in an easy to solve dataset), and provide further insights about our data. In these baselines we focus on detecting valid anchor/complement pairs, without considering the preposition’s identity.

Title Link This baseline considers one of the title’s NPs as the *complement* for each NP in the text. We experiment with three variants: *Title-First*, *Title-Last*, and *Title-Random*, which use the first, last, and a random NP in the title, respectively.

Adjacent Link The second baseline predicts the adjacent NP as a complement. We have two variants: Predict the next NP as the complement (*Adj-Forward*) or the previous NP (*Adj-Backward*).

Surface Link The third baseline predicts surface links in the text, namely, links in which the string “anchor preposition complement” appears as-is in the text. For instance, in “Adam’s father went to meet **the teacher** at *his school*” it will predict the link (*the teacher*, at, *his school*). We also experiment with *Surface-Expand*, a relaxed version that looks for the complement at a distance of up to 10 tokens following the anchor.

Combined This baseline combines the three others, using the best strategy of each one (determined based on the empirical results), and predicts

¹⁶ <https://github.com/nlpAThits/ISNotes1.0/blob/master/doc/release.annotation.scheme.pdf>.

	Model	Precision	Recall	F1
Deterministic	Human*	94.8	94.0	94.4
	Title-First	25.6	4.1	7.1
	Title-Last	29.1	4.7	8.0
	Title-Random	27.1	4.3	7.4
	Adj-Forward	21.2	3.4	5.8
	Adj-Backward	31.6	5.1	8.7
	Surface	43.5	3.3	6.2
	Surface-Expand	14.4	37.8	20.8
	Combined	15.4	44.1	22.8
	Combined-Coref	16.4	54.7	25.2
Pretrained	Decoupled-static	10.1	58.8	17.2
	Decoupled-frozen-base	9.6	55.5	16.3
	Decoupled-frozen-large	9.7	56.2	16.5
	Decoupled-base	11.8	68.5	20.1
	Decoupled-large	12.0	69.9	20.5
	Coupled-static	59.6	14.4	23.2
	Coupled-frozen-base	60.1	8.6	15.1
	Coupled-frozen-large	58.4	11.5	19.2
	Coupled-base	60.4	41.5	49.2
	Coupled-large	65.8	43.5	52.4

Table 4: Results of the deterministic baselines and neural models on the test set. We report three metrics: the precision, recall, and F1 of the overall relation predictions. The first row is an estimated human agreement on 10% of the data, and not over the entire test set, thus marked with an asterisk. Note that the first and second parts of the table are not directly comparable, since in the Deterministic results, the preposition labels is given by an oracle, whereas in the Pretrained results, it is predicted by the models.

a link whenever at least one of the used baselines is triggered. Its purpose is to increase the recall.

Combined-Coref This final baseline adds to the *Combined* predictions the gold coreference information. For each link to an NP that is part of a coreference cluster, we also add links to all other NPs in the same cluster.

7.1 Results

The deterministic baselines’ results are summarized in the first part of Table 4.

In general, the F1 scores of the ‘single’ baselines are low, ranging between 5.8 and 20.8 points, where the *Adj-Backward* baseline achieves the lowest score and the *Surface-Expand* baseline achieves the highest score. The *Combined* baseline makes use of the best strategy of each previous baseline (based on the F1 score), that is, *Title-Last*, *Adj-Backward*, and *Surface-Expand*, and reaches 22.8 F1. *Combined-Coref* extend the *Combined* baseline by adding the coreference gold data, and

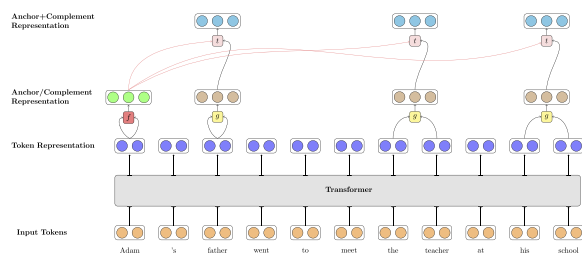


Figure 5: A schematic view of the model’s architecture.

achieves the best performance for the deterministic baselines, of an overall 25.2 F1.

These results demonstrate that (a) the links are spread across different locations in the text, and (b) that the data is unlikely to have clear shortcuts that models might exploit, while there are some strong structural cues.

8 Neural Models

Next, we experiment with three neural models based on a pre-trained masked language model (MLM), specifically, SpanBERT (Joshi et al., 2020). We also experiment with an additional baseline with uncontextualized word embeddings.

Architecture At a high level, our models take the encoding of two NPs—an anchor and a complement—and predict whether they are connected, and if so, by which preposition. To encode an anchor-complement pair, we first encode the text using the MLM and then encode each NP by concatenating the vectors of its first and last tokens. The resulting anchor and complement vectors are then each fed into a different MLP, each with a single 500-dimensions hidden-layer. The concatenation of the MLP outputs results in the anchor-complement representation. This representation is then fed into the prediction model, which has two variants. The architecture resembles the end-to-end architecture for modeling coreference resolution (Lee et al., 2017). A schematic view of the architecture is presented in Figure 5.

Variants In the **decoupled** variant, we treat each prediction as a two-step process: One binary prediction head asks “are these two NPs linked?”, and in case they are, another multiclass head determines the preposition.¹⁷ In the **coupled** variant,

¹⁷During training, the preposition-selecting head is only used for NP pairs that are connected in the gold data.

we have a single multiclass head that outputs the connecting preposition or NONE, in case the NPs are not connected. We also experiment with a **frozen** (or “probing”) variant of both models, in which we keep the MLM frozen, and update only the NP encoding and prediction heads. The frozen architecture is intended to quantify the degree to which the pretrained MLM encodes the relevant information, and it is very similar to the *edge-probing* architecture of Tenney et al. (2018). Finally, the **static** variant aims to measure how well a model can perform with NPs alone, without considering their context. This model sums all the static embeddings of each span and uses the same modeling as the coupled prediction. This baseline uses the 300-dim word2vec non-contextualized embeddings (Mikolov et al., 2013). We experiment with two versions: decoupled and coupled.

Technical Details All neural models are trained using cross-entropy loss and optimized with Adam (Kingma and Ba, 2015), using the AllenNLP library (Gardner et al., 2018). We train using a $1e^{-5}$ learning rate for 40 epochs, with early stopping based the F1 metric on the development set. We use SpanBERT (Joshi et al., 2020) as the pre-trained MLM, as it was found to work well on span-based tasks with its *base* and the *large* variants. The anchor and complement encoding MLPs have one 500-dim hidden layer and output 500-dim representations. The prediction MLPs have one 100-dim hidden layer. All MLPs use the ReLU activation. We used the same hyperparameters for all baselines and did not tune them.¹⁸

8.1 Results

In-Domain Results The pretrained models are presented in the second part of Table 4. Overall, the fully trained transformers in the *coupled* variant perform significantly better than all other models, achieving 49.2 and 52.4 F1 in the base and large variants. Interestingly, the static and frozen variants perform similarly: The F1 scores range between 15.1 and 23.2. It is worth noting that the static variant achieves better results than the frozen one. This corroborates our hypothesis that many of the capabilities needed to solve the task are not explicitly covered by the language-modeling objective and that the NPs information alone is not

¹⁸Except for the static variant for which we also tried a larger learning rate of $1e^{-3}$, which worked better in practice.

Split	Precision	Recall	F1
In-domain test	65.8	43.5	52.4
Human*	80.5	93.7	86.6
Books	46.3	28.4	35.2
Human*	89.8	89.0	89.4
IMDB	51.7	28.6	36.9
Human*	91.2	89.3	90.2
r/askedscience	48.3	25.7	33.6
r/atheism	44.1	25.7	32.5
r/LifeProTips	31.4	15.9	21.1
r/AskHistorians	36.4	26.0	30.3
r/depressed	43.4	27.3	33.5
r/YouShouldKnow	36.5	20.9	26.6
r/	37.8	22.5	28.2
Human*	86.9	90.5	88.6
OOD	46.9	27.5	34.7

Table 5: Results of the best model (Coupled-large) on OOD data, broken into the different sub-splits. The columns are the same as in Table 4. Also reporting results on in-domain split for comparison.

sufficient to solve the task, as was also argued in Hou (2020) and Pandit and Hou (2021). Finally, we note an interesting trend that the decoupled variant favors recall whereas the coupled variant favors precision, across all models. In summary, all models perform substantially below human agreement, leaving a large room for improvement.

OOD Results Here we report the best model’s results (coupled-large) on the OOD data. The results are summarized in Table 5. We break down the results per domain (and per forum in the case of Reddit), as well as the human agreement results for comparison. We observe a substantial drop in performance, with a large difference between domains (e.g., the model achieves on the IMDB split an overall 36.9 F1, while on Reddit, 28.2 F1). While the agreement scores for these domains are also lower than for the in-domain test set (88.6 F1),¹⁹ the model’s performance decreases more drastically on these splits.

¹⁹The annotation of the Wikinews domain went on for a long time, which allowed for more training, revision-and-feedback loops, and refinement of guidelines with focus on this specific type of texts and their challenges. This explains the somewhat higher agreement scores for this domain.

	Links-P	Links-R	Links-F1	IPrep-Acc	UPrep-Acc
Human*	94.8	94.0	94.4	100.0	100.0
Decoupled-static	67.3	19.6	30.4	77.7	54.2
Decoupled-frozen-base	65.9	24.5	35.7	65.0	52.5
Decoupled-frozen-large	67.2	22.8	34.1	68.2	52.7
Decoupled-base	71.1	46.7	56.4	78.2	59.9
Decoupled-large	73.5	47.2	57.5	79.3	61.5
Coupled-static	70.1	17.0	27.4	85.0	49.8
Coupled-frozen-base	73.8	10.6	18.5	81.5	44.0
Coupled-frozen-large	73.3	14.4	24.0	79.8	43.7
Coupled-base	76.4	52.4	62.2	79.1	50.7
Coupled-large	80.5	53.1	64.0	81.8	49.6

Table 6: Additional metrics of the neural models on the TNE test set. We report five metrics: the precision, recall, and F1 of the relation predictions, as well as the preposition accuracy on relations where the model predicted there is a relation (IPrep-Acc), as well as the accuracy where the model predicted there is no relation (UPrep-Acc). The first row is an estimated human agreement on 10% of the data, thus marked with an asterisk. These results are comparable with the ‘Pretrained’ part in Table 4.

9 Analysis

9.1 Quantitative Analysis

Unlabeled Accuracy and Preposition-only Accuracy To disentangle the ability to identify that a link exists between two NPs from the ability to assign the correct preposition to this link, we also report unlabeled scores (ignoring the preposition’s identity) and preposition-only scores. IPrep-Acc is the accuracy of predicting the correct preposition over gold relations (NP pairs) where the unlabeled relation was correctly identified by the model. UPrep-Acc is the accuracy of predicting the correct preposition for gold NP pairs that were not identified by the model. The results (Table 6) reveal a big gap between IPrep and UPrep accuracies for all models, indicating that the models are significantly better (yet far from perfect) at choosing the correct preposition when they identify that a relation should exist between two NPs. Overall, the preposition selection accuracy is significantly better than the majority baseline of choosing “of” for all cases (which would yield 23.5%) but substantially worse than the human agreement which is almost 100%. We also observe that while the unlabeled relation scores are indeed better than their labeled counterparts, the link-identification aspect of the task is significantly more challenging than choosing the correct preposition once the link was identified.

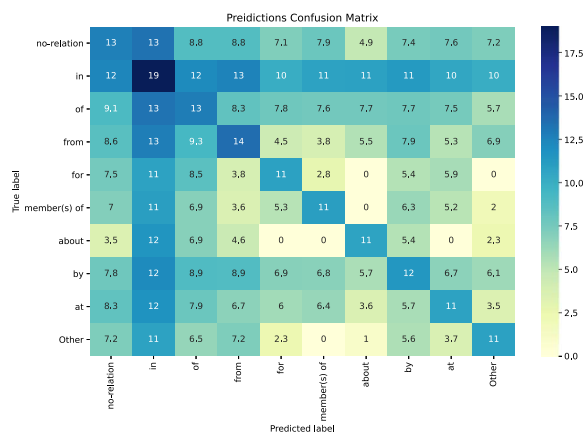


Figure 6: A confusion matrix of the predictions of the Joint-large model over the test set. The numbers are in log2 scale (except for zero values, which are untouched). We show the 10 most common labels for brevity.

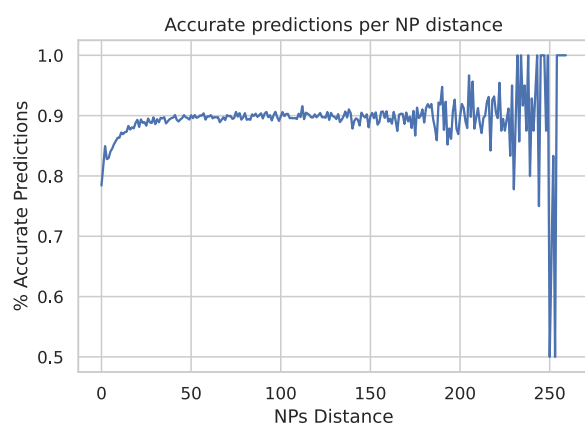


Figure 7: Accuracy of the Joint-large model over the dev set, for every NP-distance bin.

Preposition Analysis We analyze the errors of the best model on the different classes (the most common prepositions and no-relation). We present a confusion matrix in Figure 6. The most confusing label is *in*, which is confused (both in false positive and false negative) with all other labels. The preposition *of* is also confused quite frequently, while *about* is confused much less.

Accuracy per NP Distance We assess the effect of the linear distance between the two NPs on the ability of the model to accurately predict the link. For each NP pair in distance x , Figure 7 shows the percentage of correct predictions over that bin. We observe a trend of improved performance until 40 tokens, which then reaches a plateau of about 90% (the results for distances above 180 are noisy due to data sparsity at these distances). Interestingly,

the model struggles more in the short-distance links, rather than the ones farther apart. We performed the same analysis on precision and recall errors, and found similar trends.

9.2 Qualitative Analysis

To better understand the type of errors, we zoom in on a single document (shown in Figure 8 and manually inspect all errors our best model (Coupled-large) made on it.

Out of the 1980 potential links, the model wrongly predicted 231 links (82 precision errors, where a model predicted an incorrect link, and 149 recall errors, where the model failed to identify a link). Out of the 231 disagreements with the gold labels, we found 84.2% to indeed be incorrect, 10.5% to actually be correct, and 5.3% were found to be ambiguous.

Table 7 breaks down the errors into 9 categories, covering both type of errors and skills needed to solve them: *Preposition Semantics*: where the model predicted a link, but used a wrong preposition; *Ambiguous*: where both gold and predicted answers can be correct, depending on the reading of the text; *Wrong Label*: where the gold label is incorrect; *Missing Label*: where the prediction was correct, as well as the original label, but the predicted preposition was missing (i.e., cases where more than one preposition are valid); *Generics*: cases where the anchor is generic, and thus no link exists from it; *Coreference Error*: where the model links to a complement that appears to be part of the coreference chain, but is not, or the annotator mistakenly attached an additional, erroneous NP to the chain; *World Knowledge*: links that require some world knowledge in order to complete; *Explicit*: where the link appears explicitly in the text, but the model did not predict it accordingly; and *Other*, for none of the above.

In Table 4 we observed that the model is better at precision than recall. Here we also observe that recall and precision errors differ also in their type distribution. In terms of precision, 17.0% of the links were correct links with a wrong preposition. Such errors seem rather trivial, such that a good language model would not err: Using an LM for explicitly quantifying the likelihood of links may be a promising direction for future work. An interesting error that occurs both in the precision and recall errors is that of Ambiguous categorization—for instance, in the recall category, one interpretation can be read as an *opinion* being expressed

Church of Scientology does not see humor in website dedicated to Tom Cruise

September 25, 2005

<http://ScienTOMogy.info> has apparently received a fax and at least 6 emails in the span of 2 days from Scientology lawyer Ava Paquette of Moxon & Kobrin threatening a lawsuit of up to \$100,000 if the domain name ownership is not transferred. This type of letter is often called a cease and desist letter.

The owners of [scienTOMogy.info](http://ScienTOMogy.info) have posted the complaints and their replies, saying that the site simply expresses opinion, does not make any claims, and clearly states that it has no connection to the Church of Scientology. "The site was put up as a single source to view all the recent hype Tom has made about the church - it does nothing but show Tom, so we are at a loss as to why the church is acting so rashly."

The Church of Scientology is notorious for pursuing legal action against its critics, under the name of the "Religious Technology Center" (RTC). It previously made headlines when it used the US's Digital Millennium Copyright Act to remove xenu.net, a site critical of Scientology, from Google's listings.

Figure 8: Development-set document used for the qualitative error analysis. All 45 considered NPs are underlined. Out of $45^2 - 45 = 1980$ potential links, this document contains 271 gold links, and 231 erroneously predicted links, which we analyze.

	Error Type	% (Number)	Anchor	Label	Complement	Prediction
Precision Errors	Preposition Semantics	17.0 (14)	<u>a lawsuit</u>	<u>by</u>	<u>Church of Scientology</u>	<u>against</u>
	Ambiguous	14.6 (12)	<u>opinion</u>	<u>about</u>	<u>Church of Scientology</u>	<u>of</u>
	Wrong Label	13.4 (11)	<u>a fax</u>	<u>no-relation</u>	<u>the Church of Scientology</u>	<u>about</u>
	Missing Label	13.4 (11)	<u>a fax</u>	<u>to</u>	<u>the site</u>	<u>about</u>
	Generics	12.1 (10)	<u>letter</u>	<u>no-relation</u>	<u>Church of Scientology</u>	<u>to</u>
	Coreference Error	7.2 (1)	<u>at least 6 emails</u>	<u>no-relation</u>	<u>Scientology</u>	<u>about</u>
	Other	21.9 (18)	<u>their replies</u>	<u>no-relation</u>	<u>Scientology</u>	<u>to</u>
Recall Errors	World-Knowledge	12.7 (18)	<u>Church of Scientology</u>	<u>from</u>	<u>US</u>	<u>no-relation</u>
	Wrong Label	8.5 (12)	<u>the complaints</u>	<u>in</u>	<u>a fax</u>	<u>no-relation</u>
	Ambiguous	4.9 (7)	<u>opinion</u>	<u>about</u>	<u>Tom Cruise</u>	<u>no-relation</u>
	Explicit	3.5 (5)	<u>Scientology lawyer Ava Paquette</u>	<u>of</u>	<u>Moxon & Kobrin</u>	<u>no-relation</u>
	Coreference Error	2.8 (4)	<u>a fax</u>	<u>to</u>	<u>a single source</u>	<u>no-relation</u>
	Other	67.3 (95)	<u>Website</u>	<u>about</u>	<u>Tom</u>	<u>no-relation</u>

Table 7: Error types, and their statistics, based on the text presented in Figure 8. The first part of the table presents precision errors, where the model predicted some link considered to be an error. The second part presents recall errors, where the model predicted no link exists.

about *Tom Cruise*, while the other interpretation reads *opinion* in a more abstract way, thus not connected to Cruise. Finally, the largest category in the precision errors and the most common category in recall errors is, “Other”, with varied mistake that do not single out noticable phenomena.

10 Related Tasks and Linguistic Phenomena

From the outset, recovering NP-NP relations appears familiar from many previous linguistic endeavors. While TNE is related to them, it is certainly different, in scope, purpose, and definition.

Our departure point for this work has been the notion of an *implicit argument of a noun*, that is, nouns such as “brother” or “price” that are incomplete on their own, and require an argument to be complete. In linguistics, these are referred to as *relational nouns* (Partee, 1983/1997; Loebner,

1985; Barker, 1995; De Bruin and Scha, 1988; Partee et al., 2000; Löbner, 2015; Newell and Cheung, 2018). In contrast, nouns like “plant”, or “sofa” are called *sortal* and are conceived as “complete”; their denotation need not rely on the relation to other nouns, and can be fully determined.

A sensible task, then, could be to identify all the relational nouns in the text and recover their missing noun argument. However, in practical terms, the distinction between sortal and relational is not clear-cut. Specifically, sortal nouns often stand in relations to other nouns, and these relations are useful for understanding the text and for fully determining the reference—as in “the sofa [in the house]”, or “the sofa [on the carpet]” (as opposed to that on the floor), and “the house [of a particular owner]”.

A closely related linguistic concept and an established task in the last decade is *bridging*

Description/Paper	ISNotes (Markert et al., 2012)	BASHI (Rösiger, 2018a)	ARRAU (Rösiger, 2018b)	TNE (ours)
The anchor/bridging expression can be discourse-old	No	No	No	Yes
The anchor/bridging expression has to be anaphoric (not interpretable without the antecedent)	Yes	Yes	No. “Most bridging links are purely lexical bridging pairs which are not context-dependent (e.g., Europe – Spain or Tokyo – Japan).” (Hou, 2020)	No
Cataphoric links (to expressions that appear later in the text) are allowed	No	No	No	Yes
Links are annotated as part of a larger task (e.g. IS, anaphoric phenomena)	Yes	No	Yes	No
The relations in the links have to be implicit	Yes	Yes	Yes	No
The relations in the links are limited to certain semantic types	No. Any relations are allowed, but, similarly NP Enrichment, “you must be able to rephrase the bridging entity by a complete phrase including the bridging entity and the antecedent.” If the antecedent is an NP, rephrasals are restricted to a PP or possessive/Saxon genitive. “Set bridging” is allowed in special cases.	No	Yes. Bridging is limited to a set of relations (part-of, element, subset, “other”, “undersp-rel”) (Uryupina et al., 2019). On the other hand, the “undersp-rel” category can include any relations. The relations are marked.	No. Any relations that can be expressed with a preposition, are included, as well as element-set and subset-set relations.
The antecedent/complement can be not only nominal, but also verbal or clausal	Yes	Yes	No. All bridging antecedents are nominal.	No. Only nominal complements are included
The bridging expression/anchor has to be definite	No.	No, but different labels are used to distinguish definite and indefinite expressions.	No	No. Anchors can be both definite and indefinite
Multiple antecedents / complements are allowed	Yes, but only if they have different mandatory roles in the argument structure of the bridging expression.	“As a general principle, one antecedent has to be chosen. In special cases, e.g. comparative cases where two antecedents are needed, the annotator may create two or several links.” ²⁰	Multiple antecedents are not allowed by the guidelines but in practice do occur in some cases where two antecedents appeared equally strong. However, such cases are being removed from ARRAU release 3 (forthcoming).	All the complements of every anchor should be annotated. Multiple complements are allowed and very common.

Table 8: Bridging anaphora resolution vs. NP Enrichment comparison.

anaphora resolution (Clark, 1975; Loebner, 1998; Poesio and Vieira, 1998; Matsui, 2001; Gardent et al., 2003; Markert et al., 2012; Hou et al., 2013a,b; Nedoluzhko, 2013; Hou et al., 2014; Grishina, 2016; Rösiger, 2018a; Hou et al., 2018; Hou, 2018a,b, 2020; Pagel and Roesiger, 2018; Roesiger et al., 2018a; Rösiger, 2018b; Pandit and Hou, 2021; Kobayashi and Ng, 2021; Hou, 2021). Both bridging anaphora resolution and NP Enrichment relate entities mentioned in the text via non-identity relations. However, there are a number of major differences between bridging and NP Enrichment. These differences are summarized in Table 8, and expanded upon in what follows.

First, there is no agreed-upon definition of bridging (Roesiger et al., 2018b). Consequently, manual annotation of bridging relations, and the use of these annotations, requires substantial

expertise and effort. In contrast, NP Enrichment is compactly defined, and is amenable to large-scale annotation after only a brief annotator training.

Secondly, the relation between a bridging expression and its antecedent²¹ has to be implicit. In NP Enrichment the relations between the anchor and the complement are either implicit or explicit.

Next, in most bridging studies a bridge is a type of *anaphora*: The bridging expression is not interpretable without the antecedent. In NP Enrichment the anchor can in fact be interpretable on its own—the complement supplements it with additional information (“sofa [on the carpet]”) or simply exposes existing information in a uniform way.

Also, bridging expressions are not discourse-old, that is, they can only refer to entities that are mentioned in the text for the first time. This

²⁰See annotation guidelines for BASHI: <https://www.ims.uni-stuttgart.de/documents/team/alt-nicht-mehr-da/roesigia/guidelines-bridging-en.pdf>

²¹In these studies the terms *bridging expression* (or *anaphoric NP*) and *antecedent* roughly correspond to our *anchor* and *complement*, respectively.

implies that in a coreference chain only the first mention can have a bridging link. In NP Enrichment there is no such restriction: An anchor can be either old or new. Furthermore, in many bridging works the antecedent does not have to be an NP. It can be also a verb or a clause. In NP Enrichment both the anchor and the complement have to be NPs.

Finally, all the aforementioned studies have been defined by and written for linguists, using linguistic terminology, with a predominantly documentary motivation. As a result the task definitions are often narrowly scoped, highly technical, and non-interpretable for non-experts—making their annotation by crowd-workers essentially impossible. It also makes the consumption of the output by (non-linguist) NLP practitioners doubtful.

In this work we aimed to define a linguistically meaningful yet simple, properly scoped, and easy to communicate task. We want crowd-workers as well as downstream-task designers to be able to properly understand the task, its scope and its output, and we want the data collection procedure to be amenable to high inter-annotator agreement.

A Note on Decontextualization Recently, Choi et al. (2021) introduced the text-decontextualization task, in which the input is a text and an enclosing textual context, and the goal is to produce a standalone text that can be fully interpreted outside of the enclosing context. The decontextualization task involves handling multiple linguistic phenomena, and, in order to perform it well, one must essentially perform a version of the NP Enrichment task. For example, decontextualizing “Prices are expected to rise” based on “Temporary copper shortage. Prices are expected to rise”, involves establishing the relation “Prices [of copper] are expected to rise”).

Like our NP Enrichment proposal, the decontextualization task bears a strong application-motivated, user-facing perspective. It is useful, well-defined and easy to explain. However, as it is entirely goal-based (“make this sentence standalone”), the scope of covered phenomena is somewhat eclectic. More importantly, the output of the decontextualization task is targeted at human readers rather than machine readers. For example, it does not handle relations between NPs that appear within the decontextualized text itself; it only recovers relations of NPs with the sur-

rounding context. Thus, many implicit NP relations are left untreated.

11 Conclusions

We propose a new task named *Text-based NP Enrichment*, or TNE, in which we aim to annotate each NP with all its relations to other NPs in the text. This task covers a lot of implicit relations that are nonetheless crucial for text understanding. We introduce a large-scale dataset enriched with such NP links, containing 5.5K documents and over 1M links—enough for training large neural networks—and provide high-quality test sets, both in and out of domain. We propose several baselines for this task and show that it is challenging—even for state-of-the-art LM-based models—and that there is a big gap from human performance. We release the dataset, code, and models, and hope that the community will adopt this task as a standard component of the NLP pipeline.

Acknowledgments

We would like to thank the NLP-BIU lab, Nathan Schneider, and Yufang Hou for helpful discussions and comments on this paper. We also thank the anonymous reviewers and the action editors, Marie-Catherine de Marneffe and Mark Steedman, for their valuable suggestions. Yanai Elazar is grateful to be supported by the PBC fellowship for outstanding PhD candidates in Data Science and the Google PhD fellowship. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme, grant agreement no. 802774 (iEXTRACT) and grant agreement no. 677352 (NLPRO).

References

- Christian Barker. 1995. *Possessive Descriptions*, Dissertations in linguistics. Center for the Study of Language and Information. Bibliogr. pages 189–194.
- Robyn Carston. 2009. The explicit/implicit distinction in pragmatics and the limits of explicit communication. *International Review of Pragmatics*, 1:35–62. <https://doi.org/10.1163/187731009X455839>
- Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Realistic

- evaluation principles for cross-document co-reference resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.starsem-1.13>
- Pengxiang Cheng and Katrin Erk. 2019. Implicit argument prediction as reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6284–6291. <https://doi.org/10.1609/aaai.v33i01.33016284>
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461. https://doi.org/10.1162/tacl_a_00377
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.
- Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*. <https://doi.org/10.3115/980190.980237>
- Jos De Bruin and Remko Scha. 1988. The interpretation of relational nouns. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 25–32. <https://doi.org/10.3115/982023.982027>
- Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859. <https://doi.org/10.18653/v1/2020.acl-main.701>
- Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060. <https://doi.org/10.18653/v1/P18-1191>
- Gottlob Frege. 1960. On sense and reference. In Darragh Byrne and Max Kölbel, editors, *Arguing About Language*, pages 36–56. Routledge.
- Claire Gardent, Hélène Manuélian, and Eric Kow. 2003. Which bridges for bridging definite descriptions? In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-2501>
- Matthew Gerber and Joyce Y. Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798. <https://doi.org/10.1162/COLLa.00110>
- Luke Gessler, Shira Wein, and Nathan Schneider. 2021. Supersense and sensibility: Proxy tasks for semantic annotation of prepositions. In *SCIL*.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288. <https://doi.org/10.1162/089120102760275983>
- Yulia Grishina. 2016. Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 7–15, San Diego, California. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-0702>
- Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96*, pages 466–471, USA. Association for Computational Linguistics. <https://doi.org/10.3115/992628.992709>
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role

- labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653. <https://doi.org/10.18653/v1/D15-1076>
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.
- Yufang Hou. 2018a. A deterministic algorithm for bridging anaphora resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1938–1948, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1219>
- Yufang Hou. 2018b. Enhanced word representations for bridging anaphora resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 1–7, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-2001>
- Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438. <https://doi.org/10.18653/v1/2020.acl-main.132>
- Yufang Hou. 2021. End-to-end neural information status classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1377–1388. <https://doi.org/10.18653/v1/2021.findings-emnlp.119>
- Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, Washington, USA. Association for Computational Linguistics.
- Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, Georgia. Association for Computational Linguistics.
- Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093, Doha, Qatar. Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1222>
- Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284. <https://doi.org/10.1162/colia.00315>
- Marwan Jarrah. 2016. Explicit-implicit distinction: A review of related literature. *Advances in Language and Literary Studies*, 7:175–184. <https://doi.org/10.7575/aiac.all.s.v.7n.1p.175>
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. https://doi.org/10.1162/tacl_a_00300
- D. Jurafsky and J. H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall.
- Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015. <https://doi.org/10.18653/v1/D18-1546>
- Katherine Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O’Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1547–1557,

- Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1163>
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.
- Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.274>
- Hideo Kobayashi and Vincent Ng. 2021. Bridging resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1652–1659. <https://doi.org/10.18653/v1/2021.naacl-main.131>
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.
- Sebastian Löbner. 2015. Functional concepts and frames. In *Meanings, Frames, and Conceptual Representation*, page 15–42, Düsseldorf. Düsseldorf University Press.
- Sebastian Loebner. 1985. Definites. *Journal of Semantics*, 4. <https://doi.org/10.1093/jos/4.4.279>
- Sebastian Loebner. 1998. Definite associative anaphora. In *Approaches to Discourse Anaphora. Proceedings of DAARC96 – Discourse Anaphora and Resolution Colloquium*, Lancaster. Lancaster University.
- Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.
- Tomoko Matsui. 2001. Experimental pragmatics: Towards testing relevance-based predictions about anaphoric bridging inferences. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 248–260. Springer. https://doi.org/10.1007/3-540-44607-9_19
- A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. Annotating noun argument structure for Nom-Bank. In *Proceedings of LREC-2004*, Lisbon, Portugal.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111, Sofia, Bulgaria. Association for Computational Linguistics
- Edward Newell and Jackie C. K. Cheung. 2018. Constructing a lexicon of relational nouns. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Janis Pagel and Ina Roesiger. 2018. Towards bridging resolution in German: Data analysis and rule-based experiments. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 50–60, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0706>
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106. <https://doi.org/10.1162/0891201053630264>
- Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*, Synthesis Lectures on Human Language Technology

- Series, Morgan and Claypool. <https://doi.org/10.2200/S00239ED1V01Y200912HLT006>
- Onkar Arun Pandit and Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163. <https://doi.org/10.18653/v1/2021.naacl-main.327>
- Barbara Partee. 1983/1997. Uniformity vs. versatility: The genitive, a case study. Appendix to Theo Janssen, 1997. In Johan Benthem and Alice Meulen, editors, *Compositionality. The Handbook of Logic and Language*, pages 464–70. Elsevier.
- Barbara Partee, and V. Borshev. 2000. Genitives, relational nouns, and the argument-modifier distinction. *ZAS Papers in Linguistics*, 17. <https://doi.org/10.21248/zaspil.17.2000.46>
- Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.
- Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. Qadiscourse-discourse relations as qa pairs: Representation, crowdsourcing and baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819. <https://doi.org/10.18653/v1/2020.emnlp-main.224>
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. <https://doi.org/10.18653/v1/D16-1264>
- Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
- Ina Roesiger, Maximilian Köper, Kim Anh Nguyen, and Sabine Schulte im Walde. 2018a. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 44–49, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0705>
- Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018b. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.626>
- Ina Rösiger. 2018a. Bashi: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Ina Rösiger. 2018b. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–33, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0703>
- Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2016. Framenet II: Extended theory and practice, Technical report, International Computer Science Institute.
- Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer.

2009. SemEval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 106–111, Boulder, Colorado. Association for Computational Linguistics. <https://doi.org/10.3115/1621969.1621988>
- Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O’Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses. In *Proceedings of the 10th Linguistic Annotation Workshop*, pages 99–109, Berlin, Germany. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-1712>
- Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah Moeller, Aviram Stern, Adi Shalev, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196. <https://doi.org/10.18653/v1/P18-1018>
- Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA. Association for Computational Linguistics. <https://doi.org/10.3115/v1/W15-1612>
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU corpus. *Natural Language Engineering*, 26:1–34. <https://doi.org/10.1017/S1351324919000056>
- Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6–8, 1995*. <https://doi.org/10.3115/1072399.1072405>
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, and Michelle Franchini. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.