Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1110–1117
Marseille, 11–16 May 2020

WikiPossessions: Possession timeline generation as an evaluation benchmark for machine reading comprehension of long texts

Dhivya Chinnappa[†], Alexis Palmer[*], Eduardo Blanco[*]

[†]Thomson Reuters
dhivya.chinnappa@thomsonreuters.com
[*]University of North Texas
alexis.palmer@unt.edu, eduardo.blanco@unt.edu

Abstract

This paper presents **WikiPossessions**, a new benchmark corpus for the task of *temporally-oriented possession (TOP)*, or tracking objects as they change hands over time. We annotate Wikipedia articles for 90 different well-known artifacts (paintings, diamonds, and archaeological artifacts), producing 799 artifact-possessor relations with associated attributes. For each article, we also produce a full possession timeline. The full version of the task combines straightforward entity-relation extraction with complex temporal reasoning, as well as verification of textual support for the relevant types of knowledge. Specifically, to complete the full TOP task for a given article, a system must do the following: a) identify possessors; b) anchor possessors to times/events; c) identify temporal relations between each temporal anchor and the possession relation it corresponds to; d) assign certainty scores to each possessor and each temporal relation; and e) assemble individual possession events into a global possession timeline. In addition to the corpus, we release evaluation scripts and a baseline model for the task.

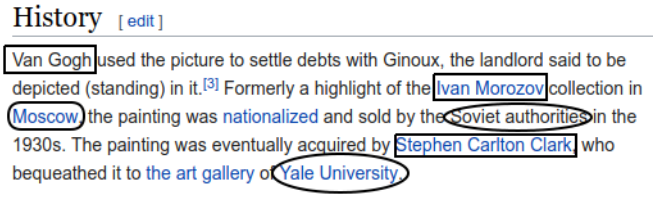
Keywords: possession, timeline, wikipossesions

1. Temporally-oriented possession

Research in machine reading, text comprehension, and natural language understanding continues to proceed at a rapid pace. Each time a new approach claims to surpass human performance on benchmark evaluation data sets, a new data set comes along, offering a new twist on the previous challenges (Bowman et al., 2015; Rajpurkar et al., 2016; Dua et al., 2019, for example). At the same time, the community has questioned whether these data sets in fact require comprehension of reading texts by the machine (Kaushik and Lipton, 2018). With this paper, we offer a new benchmark evaluation data set which combines several related tasks requiring different degrees of inferential complexity. Rather than focusing on question-answer pairs linked to particular text passages, the unit of analysis for this data set is a complete Wikipedia article. As a comparison, the average text length in our evaluation data set is an order of magnitude longer than the average passage in the DROP corpus (Dua et al., 2019), which was developed specifically to demand reasoning across longer text spans than previous reading comprehension benchmarks.

The framework for the evaluation is the task of *temporally-oriented possession (TOP)*, or tracking objects as they change hands over time. To complete the full TOP task for a given text, a system must do the following:

a) identify all possessors of a target possessee;
b) anchor possessors to times or events;
c) determine temporal relations between each temporal anchor and the possession relation it corresponds to;
d) label both possessors and temporal relations as certain or uncertain, according to whether the text provides explicit evidence; and


History [edit]

Van Gogh used the picture to settle debts with Ginoux, the landlord said to be depicted (standing) in it.[3] Formerly a highlight of the Ivan Morozov collection in Moscow, the painting was nationalized and sold by the Soviet authorities in the 1930s. The painting was eventually acquired by Stephen Carlton Clark, who bequeathed it to the art gallery of Yale University.

Figure 1: Excerpt from the Wikipedia article *The Night Café*. The possessors are highlighted. The rectangles represent persons, the curved rectangles represent locations and the ovals represent organizations.

e) assemble all individual possessors into a complete, ordered possession timeline.

The timeline generation task described above requires reasoning over the entire text, and each subtask requires different types of local information. This subtask set-up allows for breaking the input text into manageable chunks (e.g. handling one possessor at a time), making the task suitable for most state-of-the-art neural architectures.

Our contributions are a new corpus annotated for all temporally-oriented possession subtasks, an evaluation procedure, and a simple baseline model The data, evaluation scripts, and baseline model are publicly available.[1]

The TOP framework can be applied to any text which describes multiple changes of possession of one or more entities. In this work, we realize the framework in one particular domain. The corpus (Section 4.), **WikiPossessions**, is a collection of Wikipedia articles for 90 different well-known artifacts (paintings, diamonds, and archaeological artifacts), producing 799 artifact-possessor relations with

Work done at the University of North Texas

[1]Data available at dhivyachinnappa.com

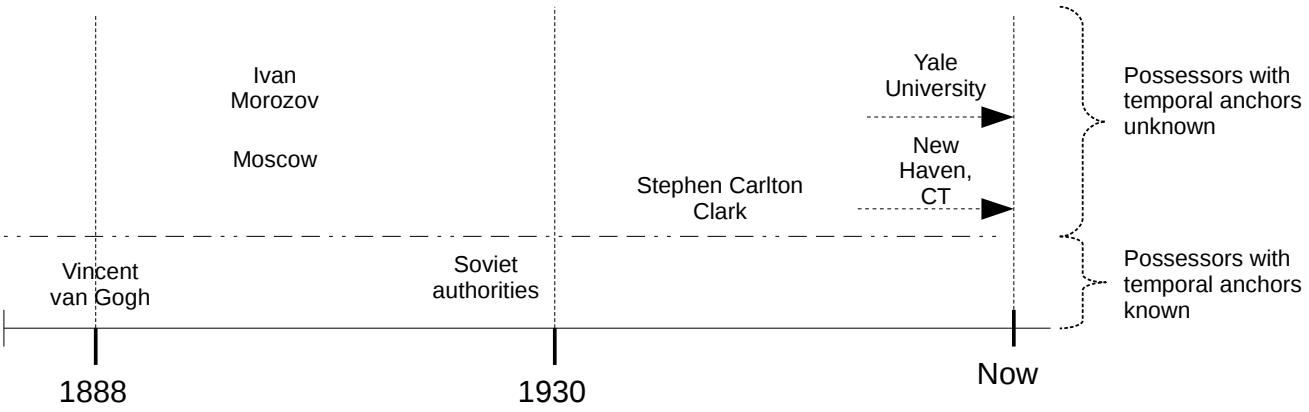


Figure 2: Possession timeline for the Wikipedia article *The Night Café* (see Fig. 1 for excerpt of text).

associated attributes. Figure 1 is an excerpt from one article in the corpus, with all possessors highlighted. For each article, we also produce a full possession timeline; Figure 2 is a visualization of the complete timeline produced for this article. All annotations are manually-validated.

Although the various subtasks each contribute to the end goal of generating a possession timeline, each subtask is evaluated independently (Section 5.). The baseline model (Section 6.) uses simple heuristics to accomplish each subtask.

2. Possession

From a linguistic perspective, the term *possession* refers to a particular set of semantic relations between two entities, the *possessor* and the *possessee* (Stassen, 2009). A wide range of different asymmetric relationships fall under the heading of possession, including kinship, proximity, part-whole relations, experience of abstract concepts, and physical possession, both permanent and temporary. The most typical notion of possession involves ownership or control of the possessee by the possessor, as in phrases like "my piano," "the lion's beautiful tail," or "this friend of mine." The linguistic literature makes a conceptual distinction between *alienable possession*, in which possessees can be separated from their possessors, and *inalienable possession*, in which such separation is not possible (Aikhenvald and Dixon, 2012; Heine, 1997, among others). Unlike inalienable possessions, which are permanent, alienable possessions are temporary and, therefore, capable of changing hands. We are interested in tracking change of possession and thus restrict our work to alienable possessions.

Previous work on automatic extraction of possession has mostly focused on particular syntactic constructions. Tratz and Hovy (2013) investigate various semantic relations realized by English possessive constructions, and both Nakov and Hearst (2013) and Tratz and Hovy (2010) consider possession expressed by noun compounds, such as "family estate." Badulescu and Moldovan (2009) extract possession as one of the many semantic relations expressed by English genetives. Blodgett and Schneider (2018) present a corpus of web reviews annotating genitives with adpositional supersenses, finding that this inventory works well for canonical possessives. We consider *all* expressions of possession, whether phrasal, clausal, sentential, or even inter-sentential. The non-restrictive approach presented here is similar to that of Banea et al. (2016), who annotate possessions of particular bloggers at the time of utterance.

In our previous work, we (Chinnappa and Blanco, 2018a) first extract possessions from a sentence using a deterministic procedure and then identify the types and temporal anchors of possession. In a latter work, we (Chinnappa and Blanco, 2018b) work with the same Wikipedia articles about artworks as the current work, and extract their possessors. We match these possessors to their temporal information with respect to the years explicitly mentioned (before, during or after). Unlike the current work, we limit the temporal information to be a year within a Wikipedia section. We do not capture any other possession attributes such as certainty or order.

3. Reasoning about possessions

The *temporally-oriented possession* task consists of several subtasks with different degrees of inferential complexity. There are three stages to the process of producing a possession timeline with all associated attributes, and each stage depends on input from one or more previous stages.

Stage One: Possessor Identification and Certainty. The first step takes as input a text passage and a target possessee; output from this step is a list of possessors—entities who have possessed the target entity at some point in time. In our case, the text passage is a Wikipedia article about a single well-known artifact (the target possessee).

Possessor identification is a fairly straightforward entity relation extraction task. What makes this data set more challenging is the requirement to extract all relations of this type from a relatively long input text. The target possessee, as the subject of the article, is often left implicit (ex. 1, where the reader must infer what is owned by Yale) or referred to pronominally (ex. 2).[2]

(1) On March 27, 2016 the United States Supreme Court rejected an appeal by Konwaloff regarding the case ... The rejection means **Yale**'s ownership is absolute. [*The Night Café*]

[2]In examples, possessors (and sometimes temporal anchors) appear in boldface, and the Wikipedia article name in square brackets.

All possessors	
Total for all articles	799
NE type: Pers / Org / Loc	281 / 318 / 200
Poss: Cert / Uncert	774 / 25
Time: Known / Unknown	**660** / 139
TRel: Before / During / After	7 / 647 / 6
TRel: Cert / Uncert	608 / 52

Table 1: Statistics for all marked possessors

Total # of Wikipedia articles	90
Total # of mentioned possessors	799
Total # of unique possessors	735
Avg # of words per article	2315
Avg # of sections per article	6.66
Avg # of possessors per article	8.87
Avg # of unique possessors per article	7.99

Table 2: Corpus statistics for WikiPossessions.

(2) It was loaned to the **Ashmolean Museum** in early 1900s, its whereabouts after this are unknown; it was rediscovered in **a Battersea home** in the early 1960s, boxed in over a chimney. [*Flaming June*]

After identifying a set of possessors, the system next should determine how certain the possession relation is *according to textual evidence* (section 4.2.2.). Certainty of possession is a property of an individual `(artifact, possessor)` relation.

Stage Two: Temporal Anchoring, Temporal Relations, and Certainty. The second step takes as input a text passage and an extracted `(artifact, possessor)` relation; output is a temporal anchor for the possession relation. This task corresponds to the TLINK task in the TempEval shared tasks (Verhagen et al., 2010; UzZaman et al., 2013), limiting the range to only possession-type events. In our case, we allow three types of anchors: a year, a range of years, or a major historical event (as in ex. 3).

(3) After **the victory of Francisco Franco in Spain**, the painting was sent to **the United States** to raise funds and support for Spanish refugees. [*Guernica*]

Once the temporal anchor has been identified, the next task is temporal relation identification (Verhagen et al., 2010; UzZaman et al., 2013). Specifically, the system should indicate whether the possession event held BEFORE the temporal anchor, DURING the temporal anchor, or AFTER (section 4.2.3.). Finally, the temporal relation is labeled with the certainty of that relation, again according to textual evidence. Temporal relation certainty is a property of an individual `(artifact, possessor, temporal anchor)` triple.

Stage Three: Possession Timeline Generation. The final stage in the process is to build a global possession timeline. The timeline should represent the order of possession between all identified possessors. The input to this final stage is the article, together with the set of extracted possession relations and all associated attributes; the output is a set of order indices, one for each `(artifact, possessor)` relation. Note that *all* identified possessors must be included in the timeline, regardless of whether a temporal anchor has been identified for the possessor. Producing this complete possession timeline thus requires more than just ordering the set of temporal anchors; possessors without explicit temporal anchors must be inserted in the order as well. Figure 2 is a visualization of the final timeline output for the Wikipedia article about Van Gogh's *The Night Café*.

4. WikiPossessions: the corpus

This section describes the data and annotation process used to create WikiPossessions.

4.1. Data: articles about famous artifacts

We collected a corpus of English Wikipedia articles about historical artifacts that could possibly change hands over time, being held by different possessors in different years. The article topics include paintings, diamonds, relics, sculptures, and archaeological findings.

Next, the set of articles was filtered to retain only articles that: a) do not discuss more than one artifact; and b) contain at least three possessors for the artifact. These filtering criteria are motivated by the goal of automatically extracting possession timelines from the texts. The resulting corpus consists of 90 articles, with each article focusing on a single target artifact. For a given article, the target artifact is the possessee in all identified possession relations. Table 2 shows basic statistics for the corpus. Note that we count both the total number of possessors and the number of unique possessors.

4.2. Annotating temporally-oriented possession

The annotation scheme was designed primarily to capture all temporal information relevant to changes of possession over time. Thus, in addition to identifying artifact-possessor relations (Section 4.2.1.), we identify a temporal anchor for each relation and the temporal relation of the possession with respect to the temporal anchor (Section 4.2.3.). The set of possession relations is then ordered into a timeline (Section 4.2.4.). For both possession relations and temporal relations, we annotate whether these features are certain or not, given the available textual evidence (Section 4.2.2.).

Annotators were provided with HTML pages of the 90 selected Wikipedia articles.[3] Annotation was done using the Wired-Marker[4] Firefox extension to annotate the HTML pages. First, all possessors of the target artifacts (Section 4.2.1.) were highlighted, using different-colored markers (provided by Wired-Marker) for different named entity types. All other annotation features (Sections 4.2.2., 4.2.3., 4.2.4.) were added to the highlighted text using Wired-Marker's notes function.

The annotators were instructed to read the entire document to decide on the possessors and the order of possession. Unless one possessor possessed the artifact at different, non-

[3] Original download date: 12th June 2017

[4] http://www.wired-marker.org/en

NE	Possessor	Poss.Cert	Ordering	Temp.Anchor	TRel	TRel.Cert
PER	Vincent van Gogh	C	1	1888	During	C
PER	Ivan Morozov	C	2	Unknown	-	-
LOC	Moscow	C	2	Unknown	-	
ORG	Soviet authorities	C	3	1930	Before	C
PER	Stephen Carlton Clark	C	4	Unknown	-	-
ORG	Yale University	C	5	Unknown-Now	During	C
LOC	New Haven, CT	C	5	Unknown-Now	During	C

Table 3: Complete annotation, including **possession timeline**, for Wikipedia article on Van Gogh's *The Night Café*.

contiguous points in time, only one mention of each possessor is annotated.

4.2.1. Possessors and artifacts

We focus on a single possessee/artifact, namely the topic of the article. For that artifact, we identify all possessors *of that artifact* mentioned over the course of the article. The corpus consists of all artifact-possessor pairs identified from the selected Wikipedia articles. We extract 799 pairs in all, with 735 unique artifact-possessor pairs.
For example, the text snippet seen in Figure 1 yields the following pairs:[5]

```
1. (night_cafe, ivan_morozov)
2. (night_cafe, soviet_authorities)
3. (night_cafe, stephen_carlton_clark)
4. (night_cafe, yale_university)
```

The possessors identified each fall into one of three named entity (NE) categories: Person, Organization (e.g. museums or universities), or Location (e.g. particular cities, states, or countries). The NE type of each possessor is labeled manually, with the resulting distribution shown in Table 1. Organizations are the most frequent possessors, followed by People and then Locations. Although the possessors fall neatly into traditional NE categories, many of them are not in fact recognized by standard NE taggers. These include cases like example (4):

(4) On the morning of March 18, 1990, **thieves** disguised as police officers broke into the museum and stole *The Storm on the Sea of Galilee* and 12 other works. [*The Storm on the Sea of Galilee*]

The thieves who stole the painting, and presumably possessed it for at least some time thereafter, are unnamed. English NER systems also struggle to recognize possessors such as "artist's daughter" or names in other languages. This means that NER alone is not sufficient to identify possessors, even in this specific context where all possessees are concrete artifacts and likely to be owned by NEs.

4.2.2. Certainty of possession

For each artifact-possessor pair, annotators are asked to assess the certainty of the possession relation. We are interested in the notion of certainty as it relates to textual evidence: if the text *of the entire article* strongly supports the relation, the instance should be marked as Certain (C). If not, it should be marked as Uncertain (UC).
Example (5) illustrates a case of uncertainty; the phrase "generally accepted" indicates some degree of uncertainty on the part of the author.

(5) It was completed after Giorgione's death in 1510, with the landscape and sky generally accepted to have been completed by **Titian**. [*The Sleeping Venus*]

4.2.3. Temporal anchor and relations

This section describes annotation of temporal features of the extracted possession relations.

Temporal anchor. The first time-related annotation decision is to determine whether, according to the text, there is a temporal anchor for the given possession relation. For cases when a possessor has held an artifact for more than one time period, different temporal anchors may be associated with the same artifact-possessor pair, as in example (6) below. This painting was in the custody of its owner prior to the 1873 Exhibition, and then again for a period between the end of that exhibition and the painting's 1878 journey out of Russia.

(6) Despite its progressive implications, Barge Haulers was bought by the **Tsar's second son**. It was lent for exhibition at the 1873 **International Exhibition** in **Vienna**, where it won a bronze medal. It was exhibited **outside Russia** again in 1878... [*Barge Haulers on the Volga*]

If a temporal anchor cannot be identified, the other temporal features are not relevant. Looking again at Figure 1, only one of the five possessors has an identifiable temporal anchor: **Soviet authorities**, anchored in **1930**. In the corpus, 660 of 799 possessors are associated with a temporal anchor.

Temporal relations. Our temporal anchors are similar in nature to the TLINK annotations of the TimeBank Corpus (Pustejovsky et al., 2003), but we restrict the granularity to the level of the year. An anchor could denote a single year, a range of years, or some historical event. Particular days or months are ignored.
In the ideal case, the temporal anchor covers the entire period of possession (e.g. "1983-1987"), but more often the text mentions a date or historical event (e.g. World War II) which may or may not lie within the duration of possession. To build possession timelines, we need to know how the temporal anchor relates to the period of possession. Thus we annotate three different categories of temporal relation.

[5]The complete annotations for the article containing the text snippet are shown in Table 3.

BEFORE indicates that possession occurred prior to the anchor, while AFTER indicates that possession occurred later than the temporal anchor. DURING indicates that the period of possession includes the temporal anchor. Annotators also mark whether the temporal relations are certain or uncertain, according to the evidence in the text.

(7) BEFORE: At some undetermined point before **1516** it came into the possession of **Don Diego de Guevara** ... [*Arnolfini Portrait*]

(8) DURING: In **1599** a German visitor saw it in the **Alcazar Palace** in **Madrid**. [*Arnolfini Portrait*]

(9) AFTER: In **1530** the painting was inherited by Margaret's niece **Mary** of Hungary, who in 1556 went to live in Spain. [*Arnolfini Portrait*]

Note that the temporal annotations reflect only the knowledge contained in the text; they do not provide complete information about changes in possession. The temporal anchor in example (7) provides the latest possible date at which possession of the portrait transferred to Don Diego de Guevara. Example (8) conveys that the temporal anchor 1599 occurs sometime during the Alcazar's possession of the portrait. We do not know where in the period of possession the date falls; it could be a beginning or end date. The temporal anchor in example (9) marks the change of possession.

4.2.4. Possession ordering and timeline

The final annotation task is to order the possessors according to when each had control of the artifact in question, building up a **possession timeline**. Each possessor is given a serial number, depending on the order in which the artifact was possessed. Usually the artist who created or found the artifact (if known) is assigned the serial number 1. The next possessor gets the serial number 2, and so forth. An example timeline can be seen in Table 3.

Annotation relies on the complete textual context, often allowing the annotator to determine ordering of possession events even if explicit temporal anchors are not stated in the text. For example, in Figure 1, we can infer that the possessor (**Yale University**), to whom the painting was bequeathed, appears in the timeline later than the possessor who did the bequeathing.

When two possessors held the artifact simultaneously (e.g. **Yale University** and **New Haven, Connecticut**), both appear at the same position in the timeline. A possessor can receive multiple serial numbers when there are multiple relevant periods of possession (as in (6)).

4.2.5. Inter-annotator agreement

Temporally-oriented possession is a new task, with a new annotation scheme. To measure the soundness and reliability of the annotation scheme, a portion of the articles (12/90, or 13%) are annotated by a second annotator. The articles were selected randomly from a subset of articles of roughly average length for the corpus, in order to reduce oddities due to overly long (or short) texts. Inter-annotator agreement was calculated for each annotation subtask. Overall, annotator agreement is very high, suggesting that the task is well-defined and the annotation scheme reliable.

First, we look at identification of artifact-possessor pairs. We treat Annotator A's labels as a pseudo-gold standard and measure precision and recall of Annotator B's labels as compared to Annotator A. Precision is 0.97, and recall is 0.69.

Inter-annotator agreement for the temporal and certainty features is calculated only for the set of artifact-possessor pairs identified by both annotators. For both possession certainty and temporal relation certainty, Cohen's κ is very high (0.92). Agreement is more moderate, but still substantial, for the temporal features. Cohen's κ for the temporal anchor is 0.77. For temporal relations (before/during/after), Cohen's κ is 0.76. For the order of possession (the possession timeline), we generate a list of ordered pairs of possessors for both annotators and then compare. Precision between the two lists of pairs is 0.93, and recall is 0.90.

5. Evaluation

Given the gold standard annotations (WikiPossessions corpus), and a system that aims to predict those same annotations, our goal is to evaluate the performance of the system. The title of the Wikipedia article is the possessee and is fixed. Possessors are found across the entire article. The possessor is the entity of interest, and all other annotations, including possessor certainty, temporal anchor, temporal relation, temporal relation certainty, and order, are attributes of the possessor.

For each stage in the timeline generation process, we consider both a strict (**exact match**) evaluation setting and a more relaxed **partial match** setting. The types of information extracted and/or predicted allow for straightforward evaluation: the various subtasks require named entities, years (or major historical events, realized as short text spans), labels (binary for certainty tasks, 3-way for temporal relations), and ordering indices (integers). Unless otherwise indicated, we evaluate using precision, recall, and F1.

Stage One: Possessor Identification and Certainty. In the exact match setting, an extracted possessor matches gold only if the spans are an exact match. Mismatches caused by articles or prepositions on either side (gold or predicted) are ignored. An extracted possessor is considered a partial match if at least one content word (defined loosely as neither an article nor a preposition) matches between the two spans. For example, the gold span *the Netherlands* and extracted span *Netherlands* are considered an exact match; the gold span *Museum of Modern Art* and extracted span *Modern Art* are considered a partial match; *Museum of Modern Art* and *Queen of Sheba* are not considered a match.

Possessor certainty is a binary task, with possible labels Certain and Uncertain. To evaluate possessor certainty, we simply compare the predicted label to the gold label. In the exact match setting, both the certainty label and the possessor must match with the gold, in order to be considered correct. In the partial match setting, the possessor can be a partial match with gold.

	ALL				ALLCORR				ORACLE	
	Exact match		Partial match		Exact match		Partial match			
	P / R	F1	P / R	F1	P / R	F1	P / R	F1	P / R	F1
Poss.ID	0.03/0.35	0.06	0.06/0.65	0.11	1.00/0.35	0.52	1.00/0.65	0.79	1.00/0.89	0.94
Poss.Cert	0.03/0.34	0.06	0.06/0.64	0.11	0.89/0.34	0.49	0.97/0.64	0.77	0.93/0.86	0.89
Temp.Anch	0.003/0.04	0.006	0.008/0.10	0.014	0.08/0.036	0.05	0.15/0.10	0.12	0.21/0.20	0.20
Temp.Rel	0.004/0.06	0.008	0.02/0.24	0.04	0.14/0.063	0.09	0.35/0.24	0.28	0.50/0.48	0.49
TRel.Cert	0.004/0.06	0.008	0.02/0.22	0.04	0.14/0.06	0.08	0.33/0.02	0.04	0.50/0.52	0.51
Ordering	0.001/0.06	0.001	0.002/0.26	0.004	0.47/0.06	0.11	0.64/0.26	0.37	0.16/0.48	0.24

Table 4: Results for baseline model on full WikiPossessions corpus. See text for details.

Stage Two: Temporal Anchoring, Temporal Relations, and Certainty. For all three subtasks of Stage Two, exact match and partial match settings are as described for possessor certainty.

Temporal anchor evaluation must handle three different situations, with respect to the type of answer required: a) year(s) vs. year(s); b) event vs. event; and c) year(s) vs. event. Type (c) will always be considered incorrect. To evaluate type (b), we compare text spans, requiring exact match as defined above. To evaluate type (a), we first convert each date to both an upper bound and a lower bound. For example *1832* is converted to the pair `lower_bound=1832` and `upper_bound=1832`. The range *1815-1845* in converted to the pair `lower_bound=1815` and `upper_bound=1845`. Next, we generate all years between the lower and upper bounds, for both predicted temporal anchor and gold temporal anchor, and calculate precision and recall between the two sets of years.

This approach to evaluation fails to capture the importance of temporal proximity in evaluating temporal reasoning systems. For example, if the gold temporal anchor is 1832, we'd prefer a system that predicts 1834 to a system that predicts 1934, though the previously-outlined evaluation would score the two equally. To account for this, we introduce a degree of tolerance for evaluating temporal anchors. If the degree of tolerance is 25%, we allow for 25% error in predicting the temporal anchor. Specifically, we find the difference between the lower bound and the upper bound in the gold standard (x) and compute an adjustment rate y by multiplying x by the specified degree of tolerance. Next, both upper and lower bound are adjusted by y (minimum value of 1), expanding the number of years in the gold standard set.

Evaluation of temporal relations and certainty of temporal relation is relevant only for possessors with an identified temporal anchor. In both cases, we simply compare predicted labels (Before/During/After or Cert/Uncert) to gold labels, following the definitions of exact and partial match given above.

Stage Three: Possession Timeline Generation. Finally, we evaluate the possession timeline. Although the timeline annotations take the form of order indices assigned to individual possessors, comparison of indices between predicted and gold timelines would magnify errors by overly punishing a system for missing one possessor in the timeline. To give credit for correct ordering decisions, we use a pairwise evaluation strategy for possessor ordering. Pairs of possessors are generated according to their order, where the first member of the pair has an order index less than or equal to the second member of the pair, e.g. the order annotations in Table 3 would produce the following pairs (partial set):

```
1. (vincent_van_gogh, ivan_morozov)
2. (vincent_van_gogh, moscow)
3. (ivan_morozov, moscow)
4. (moscow, ivan_morozov)
5. (vincent_van_gogh, soviet_authorities)
6. (ivan_morozov, soviet_authorities)
7. (moscow, soviet_authorities)
...
```

Precision and recall are computed by comparing the two sets of pairs. In the exact match setting, both possessors in a pair must satisfy exact match conditions; for the partial match setting, both possessors must be at least a partial match with gold.

Subtask	Heuristic
Poss.ID	Extract all NEs of type PERS, ORG, LOC
Poss.Cert	Label all possessors Certain
Temp.Anch	Select DATE closest in text to possessor
Temp.Rel	Label all relations During
TRel.Cert	Label all relations Certain
Ordering	Use order of appearance in text

Table 5: Heuristics used in baseline model.

6. Baseline model

To demonstrate the viability of the TOP task, and also provide a comparison system, we implement a very simple baseline model, with no learning.

6.1. Model

The baseline model treats each subtask of the TOP task using a straightforward heuristic (Table 5). First, the data is preprocessed using spaCy (Honnibal and Montani, 2017) for tokenization, part-of-speech tagging, and named entity recognition and labeling. The tasks of possessor identification and temporal anchor extraction rely on spaCy's named entity output. All entities labeled as person, organization, or location are extracted as possessors. To extract the temporal anchor for a given possessor, the system looks for the

closest expression labeled as a date by spaCy;[6] this can occur either before or after the possessor in the text. Degree of tolerance is set at 10%, but does not improve results over a setting with no degree of tolerance.

For certainty and temporal relation identification, the baseline applies the most frequent label from the WikiPossessions corpus. To arrange possessors into a timeline, the baseline follows linear order in the text, allowing no more than one occurrence of a given possessor, no matter how often that possessor is mentioned in the original text.

6.2. Results

We evaluate the performance of our baseline model on the complete WikiPossessions corpus, testing three different scenarios. In each scenario we evaluate using a different set of extracted possessors, incorporating different amounts of knowledge. ALL indicates that we keep every possessor extracted by the baseline model using the heuristic described above. In ALLCORR(ect), we filter the possessors in ALL, keeping only those which match the gold standard. In the ORACLE scenario, we skip the baseline possessor identification step and instead use the full set of possessors from the gold standard.[7] For the first two scenarios, we compare the exact match and partial match settings, as described in Section 5.. All possessors in ORACLE are exact matches. Results appear in Table 4.

As a whole, results from the baseline system demonstrate the need for learning. The low precision of possessor identification using the baseline heuristic is expected, but less expected is the relatively low recall (0.35 exact match, 0.65 partial match). Even though the annotation instructions restrict possessors to be persons, locations, or organizations, NE recognition systems capture fewer than 75% of possessors, even in the relaxed partial match setting.

Precision increases in the ALLCORR scenario, as we filter extracted possessors by matching to the gold standard. The low recall, however, results in low performance in downstream tasks.

Finally, in the ORACLE scenario we assume the gold-standard set of possessors. The results for temporal anchor extraction are very low (0.20 F1, with similarly low precision and recall), demonstrating that the date in the text nearest to the string indicating the possessor is rarely an appropriate temporal anchor for the associated possession event. A similar conclusion can be drawn from ORACLE results for possessor ordering. Linear ordering according to appearance in the text results in poor performance; reasoning is required.

7. Conclusion

We present WikiPossessions, a new corpus of articles annotated for subtasks leading to the generation of possession timelines. The corpus is intended as a new evaluation benchmark for reading comprehension systems; successful completion of the subtasks requires inference over long text spans. Annotated data, evaluation scripts, and a heuristic baseline model will be made available.

The corpus focuses on (mostly) famous paintings, but the task is open-domain. The task formulation and evaluation methods can be used to track possession in other types of texts, such as ball possession in various sports, or even changes of employment. Temporally anchored possessions may be useful for analyzing the history of artifacts, or for enriching more general event timelines.

[6] spaCy tags years, months, events, and year ranges (e.g. 1530-1545) as dates.

[7] Recall is below 1 because, unlike the gold, the oracle doesn't allow multiple occurrences of the same possessor.

8. Bibliographical References

Aikhenvald, A. and Dixon, R. (2012). *Possession and Ownership: A Cross-Linguistic Typology*. Explorations in Linguistic Typology. OUP Oxford.

Badulescu, A. and Moldovan, D. (2009). A semantic scattering model for the automatic interpretation of english genitives. *Natural Language Engineering*, 15(2):215–239.

Banea, C., Chen, X., and Mihalcea, R. (2016). Building a dataset for possessions identification in text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Blodgett, A. and Schneider, N. (2018). Semantic supersenses for English possessives. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May. European Language Resource Association.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Chinnappa, D. and Blanco, E. (2018a). Mining possessions: Existence, type and temporal anchors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 496–505.

Chinnappa, D. and Blanco, E. (2018b). Possessors change over time: A case study with artworks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2278–2287, Brussels, Belgium, October-November. Association for Computational Linguistics.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. (2019). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *CoRR*, abs/1903.00161.

Heine, B. (1997). *Possession: Cognitive Sources, Forces, and Grammaticalization*. Cambridge Studies in Linguistics. Cambridge University Press.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Kaushik, D. and Lipton, Z. C. (2018). How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Nakov, P. I. and Hearst, M. A. (2013). Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Trans. Speech Lang. Process.*, 10(3):13:1–13:51, July.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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.

Stassen, L. (2009). *Predicative Possession*. Oxford Studies in Typology and Linguistic Theory. OUP Oxford.

Tratz, S. and Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 678–687, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tratz, S. and Hovy, E. H. (2013). Automatic interpretation of the english possessive. In *ACL (1)*, pages 372–381. The Association for Computer Linguistics.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.