Where's My Head? Definition, Data Set, and Models for Numeric Fused-Head Identification and Resolution

Yanai Elazar[†] and **Yoav Goldberg**[†*]
[†]Computer Science Department, Bar-Ilan University, Israel
[*]Allen Institute for Artificial Intelligence
{yanaiela,yoav.goldberg}@gmail.com

Abstract

We provide the first computational treatment of fused-heads constructions (FHs), focusing on the numeric fused-heads (NFHs). FHs constructions are noun phrases in which the head noun is missing and is said to be ''fused'' with its dependent modifier. This missing information is implicit and is important for sentence understanding. The missing references are easily filled in by humans but pose a challenge for computational models. We formulate the handling of FHs as a two stages process: **Identification** of the FH construction and **resolution** of the missing head. We explore the NFH phenomena in large corpora of English text and create (1) a data set and a highly accurate method for NFH identification; (2) a 10k examples (1 M tokens) crowd-sourced data set of NFH resolution; and (3) a neural baseline for the NFH resolution task. We release our code and data set, to foster further research into this challenging problem.

1 Introduction

Many elements in language are not stated explicitly but need to be inferred from the text. This is especially true in spoken language but also holds for written text. Identifying the missing information and filling in the gap is a crucial part of language understanding. Consider the sentences below:

(1) I'm **42** __, Cercie.

(2) It's worth about **two million** __.

(3) I've got two *months* left, **three** __ at the most.

(4) I make an amazing *Chicken Cordon Bleu*. She said she'd never had **one**.

In Example (1), it is clear that the sentence refers to the *age* of the speaker, but this is not stated explicitly in the sentence. Similarly, in Example (2) the speaker discusses the worth of an object in some *currency*. In Example (3), the number refers back to an object already mentioned before—*months*.

All of these examples are of *numeric fused heads* (NFHs), a linguistic construction that is a subclass of the more general *fused heads* (FHs) construction, limited to numbers. FHs are noun phrases (NPs) in which the head noun is missing and is said to be ''fused'' with its dependent modifier (Huddleston and Pullum, 2002). In the examples above, the numbers '*42*', '*two million*', '*three*', and '*one*' function as FHs, whereas their actual heads (YEARS OLD, DOLLAR, *months*, *Chicken Cordon Bleu*) are missing and need to be inferred.

Although we focus on NFHs, FHs in general can occur also with other categories, such as determiners and adjectives. For example, in the following sentences:

(5) Only the **rich** __ will benefit.

(6) I need some *screws* but can't find **any**__.

the adjective '*rich*' refers to rich PEOPLE and the determiner '*any*' refers to *screws*. In this work we focus on the *numeric fused head*.

Such sentences often arise in dialog situations as well as other genres. Numeric expressions play an important role in various tasks, including textual entailment (Lev et al., 2004; Dagan et al., 2013), solving arithmetic problems (Roy and Roth, 2015), numeric reasoning (Roy et al., 2015; Trask et al., 2018), and language modeling (Spithourakis and Riedel, 2018).

While the inferences required for NFH construction may seem trivial for a human hearer, they are for the most part not explicitly addressed by current natural language processing systems.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 519–535, 2019. https://doi.org/10.1162/TACL_a_00280
Action Editor: Yuji Matsumoto. Submission batch: 12/2018; Revision batch: 1/2019; Published 9/2019.

Index	Text	Missing Head
i	Maybe I can teach the kid a *thing* or **two** __.	*thing*
ii	you see like **3** __ or 4 *brothers* talkin'	*brothers*
iii	When the clock strikes **one**. . . the Ghost of Christmas Past	O'CLOCK
iv	My manager says I'm a perfect **10**!	SCORE
v	See, that's **one** __ of the *reasons* I love you	*reasons*
vi	Are you **two** done with that helium?	PEOPLE
vii	No **one** cares, dear.	PEOPLE
viii	*Men* are like *busses*: If you miss **one** __, you can be sure there'll be soon another **one** __. . .	*Men* | *busses*
ix	I'd like to wish a happy **1969** to our new President.	YEAR
x	I probably feel worse than Demi Moore did when she turned **50**.	AGE
xi	How much was it? **Two hundred**, but I'll tell him it's fifty. He doesn't care about the gift;	CURRENCY
xii	Have you ever had an *unexpressed thought*? I'm having **one** __ now.	*unexpressed thought*
xiii	It's a curious thing, the death of a loved **one**.	PEOPLE
xiv	I've taken **two** __ over. Some *fussy old maid* and some *flashy young man*.	*fussy old maid* & *flahy young man*
xv	[non-NFH] **One** *thing* to be said about traveling by stage.	-
xvi	[non-NFH] After **seven** long *years*. . .	-

Table 1: Examples of NFHs. The *anchors* are marked in **bold**, the heads are marked in *italic*. The missing heads in the last column are written in italic for *Reference* cases and in upper case for the *Implicit* cases. The last two rows contain examples with regular numbers—which are not considered NFHs.

Indeed, tasks such as information extraction, machine translation, question answering, and others could greatly benefit from recovering such implicit knowledge prior to (or in conjunction with) running the model.[1]

We find NFHs particularly interesting to model: They are common (Section 2), easy to understand and resolve by humans (Section 5), important for language understanding, not handled by current systems (Section 7), and hard for current methods to resolve (Section 6).

The main contributions of this work are as follows.

- We provide an account of NFH constructions and their distribution in a large corpus of English dialogues, where they account for 41.2% of the numbers. We similarly quantify the prevalence of NFHs in other textual genres, showing that they account for between 22.2% and 37.5% of the mentioned numbers.

- We formulate FH *identification* (identifying cases that need to be resolved) and *resolution* (inferring the missing head) tasks.

- We create an annotated corpus for NFH *identification* and show that the task can be automatically solved with high accuracy.

- We create a 900,000-token annotated corpus for NFH *resolution*, comprising ∼10K NFH examples, and present a strong baseline model for tackling the *resolution* task.

2 Numeric Fused Heads

Throughout the paper, we refer to the visible number in the FH as the *anchor* and to the missing head as the *head*.

In FH constructions the implicit heads are missing and are said to be *fused* with the anchors, which are either determiners or modifiers. In the case of NFH, the modifier role is realized as a number (see examples in Table 1). The anchors then function both as the determiner/modifier and as the head—the parent and the other modifiers of the original head are syntactically attached to the anchor. For example, in Figure 1 the phrase *the remaining 100 million* contains an NFH construction with the anchor *100 million*, which is attached to the sentence through the dotted black dependency edges. The missing head, *murders*, appears in red together with its missing dependency edges.[2]

[1]To give an example from information extraction, consider a system based on syntactic patterns that needs to handle the sentence ''Carnival is expanding its ships business, with 12 to start operating next July.'' In the context of MT, Google Translate currently translates the English sentence ''I'm in the center lane, going about 60, and I have no choice'' into French as ''Je suis dans la voie du centre, environ 60 ans, et je n'ai pas le choix'', changing the implicit *speed* to an explicit *time period*.

[2]An IE or QA system trying to extract or answer information about the number of murders being solved will have a much easier time when implicit information would be stated explicitly.

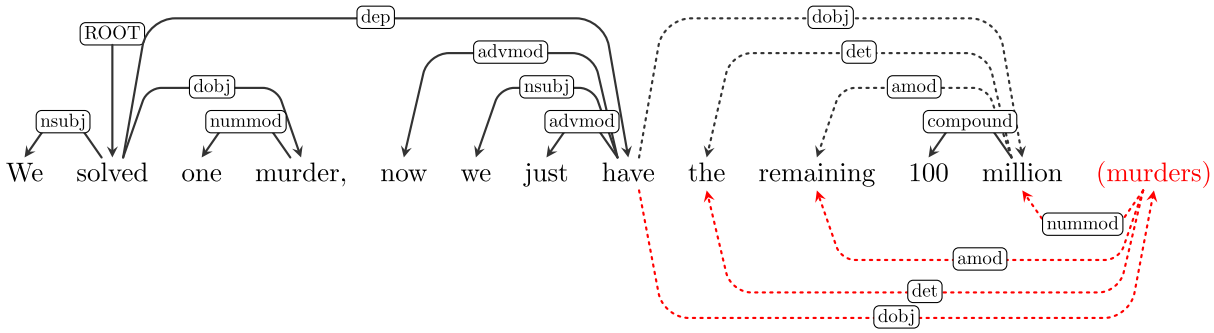

Figure 1: Example for an NFH. The 'murders' token is missing, and fused with the '100 million' numeric-span.

Distribution NFH constructions are very common in dialog situations (indeed, we show in Section 4 that they account for over 40% of the numbers in a large English corpus of movie dialogs), but are also common in written text such as product reviews or journalistic text. Using an NFH identification model that we describe in Section 4.2, we examined the distribution of NFH in different corpora and domains. Specifically, we examined monologues (TED talks; Cettolo et al., 2012), Wikipedia (WikiText-2 and WikiText-103; Merity et al., 2016), journalistic text (PTB: Marcus et al., 1993), and product reviews (Amazon reviews[3]) in which we found that more than 35.5%, 33.2%, 32.9%, 22.2%, and 37.5% of the numbers, respectively, are NFHs.

FH Types We distinguish between two kinds of FH, which we call *Reference* and *Implicit*. In *Reference* FHs, the missing head is referenced explicitly somewhere else in the discourse, either in the same sentence or in surrounding sentences. In *Implicit* FHs, the missing head does not appear in the text and needs to be inferred by the reader or hearer based on the context or world knowledge.

2.1 FH vs. Other Phenomena

FH constructions are closely related to ellipsis constructions and are also reminiscent of coreference resolution and other anaphora tasks.

FH vs. Ellipsis With respect to ellipsis, some of the NFH cases we consider can be analyzed as nominal ellipsis (cf. i, ii in Table 1, and Example (3) in the Introduction). Other cases of head-less numbers do not traditionally admit an ellipsis analysis. We do not distinguish between the cases and consider all head-less number cases as NFHs.

[3]https://www.kaggle.com/bittlingmayer/amazonreviews

FH vs. Coreference With respect to coreference, some *Reference* FH cases may seem similar to coreference cases. However, we stress that these are two different phenomena: In coreference, the mention and its antecedent both refer to the same entity, whereas the NFH anchor and its head-reference—like in ellipsis—may share a symbol but do not refer to the same entity. Existing coreference resolution data sets do consider some FH cases, but not in a systematic way. They are also restricted to cases where the antecedent appears in the discourse (i.e., they do not cover any of the NFH *Implicit* cases).

FH vs. Anaphora Anaphora is another similar phenomenon. As opposed to coreference, anaphora (and cataphora, which are cases with a forward rather than a backward reference) includes mentions of the same type but different entities. However, the anaphora does not cover our *Implicit* NFH cases, which are not anaphoric but refer to some external context or world knowledge. We note that anaphora/cataphora is a very broad concept, which encompasses many different subcases of specific anaphoric relations. There is some overlap between some of these cases and the FH constructions.

Pronimial *one* The word *one* is a very common NFH anchor (61% of the occurrences in our corpus), and can be used either as a number (viii) or as a pronoun (xiii). The pronoun usage can be replaced with *someone*. For consistency, we consider the pronominal usages to be NFH, with the implicit head PEOPLE.[4]

The *one-anaphora* phenomenon was previously studied on its own (Gardiner, 2003; Ng et al.,

[4]Although the overwhelming majority of 'one' with an implicit PEOPLE head are indeed pronomial, some cases are not. For example: '*Bailey, if you don't hate me by now you're a minority of* ***one****.*'

2005). The work by Ng et al. (2005) divided uses of *one* into six categories: Numeric (xv), Partitive (v), Anaphoric (xii), Generic (vii), Idiomatic (xiii) and Unclassified. We consider all of these, except the Numeric category, as NFH constructions.

2.2 Inclusive Definition of NFH

Although our work is motivated by the linguistic definition of FH, we take a pragramatic approach in which we do not determine the scope of the NFH task based on fine-grained linguistic distinctions. Rather, we take an inclusive approach that is motivated by considering the end-user of an NFH resolution system who we imagine is interested in resolving all numbers that are missing a nominal head. Therefore, we consider all cases that ''look like an NFH'' as NFH, even if the actual linguistic analysis would label them as gapping, ellipsis, anaphoric pronominal-one, or other phenomena. We believe this makes the task more consistent and easier to understand to end users, annotators, and model developers.

3 Computational Modeling and Underlying Corpus

We treat the computational handling of FHs as two related tasks: Identification and resolution. We create annotated NFH corpora for both.

Underlying Corpus As the FH phenomenon is prevalent in dialog situations, we base our corpus on dialog excerpts from movies and TV-series scripts (the IMDB corpus). The corpus contains 117,823 different episodes and movies. Every such item may contain several scenes, with an average of 6.9 scenes per item. Every scene may contain several speaker turns, each of which may span several sentences. The average number of turns per scene is 3.0. The majority of the scenes have at least two participants. Some of the utterances refer to the global movie context.[5]

NFH Identification In the identification stage, we seek NFH anchors within headless NPs that contain a number. More concretely, given a sentence, we seek a list of spans corresponding to all of the anchors within it. An NFH anchor is restricted to a single number, but not a single token. For example, *thirty six* is a two-token number that can serve as an NFH anchor. We assume all anchors are contiguous spans. The identification task can be reduced to a binary decision, categorizing each numeric span in the sentence as FH/not-FH.

NFH Resolution The resolution task resolves an NFH anchor to its missing head. Concretely, given a text fragment $w_1, \ldots, w_n$ (a *context*) and an NFH anchor $a = (i, j)$ within it, we seek the head(s) of the anchor.

For *Implicit* FH, the head can be any arbitrary expression. Although our annotated corpus supports this (Section 5), in practice our modeling (Section 6) as well as the annotation procedure favor selecting one out of five prominent categories or the OTHER category.

For *Reference* FH, the head is selected from the text fragment. In principle a head can span multiple tokens (e.g., 'unexpected thought' in (Table 1, xii)). This is also supported by our annotation procedure. In practice, we take the syntactic head of the multi-token answer to be the single-token missing element, and defer the boundary resolution to future work.

In cases where multiple heads are possible for the same anchor (e.g., viii, xiv in Table 1), all should be recovered. Hence, the resolution task is a function from a (text, anchor) pair to a list of heads, where each head is either a single token in the text or an arbitrary expression.

4 Numeric Fused-Head Identification

The FH task is composed of two sub-tasks. In this section, we describe the first __: identifying NFH *anchors* in a sentence. We begin with a rule-based method, based on the FH definition. We then proceed to a learning-based model, which achieves better results.

Test set We create a test set for assessing the identification methods by randomly collecting 500 dialog fragments with numbers, and labeling each number as NFH or not NFH. We observe that **more than 41% of the test-set numbers are FHs**, strengthening the motivation for dealing with the NFH phenomena.

4.1 Rule-based Identification

FHs are defined as NPs in which the head is fused with a dependent element, resulting in an

[5]Referring to a broader context is not restricted to movie-based dialogues. For example, online product reviews contain examples such as ''. . . *I had three in total...*'', with *three* referring to the purchased product, which is not explicitly mentioned in the review.

NP without a noun.[6] With access to an oracle constituency tree, NFHs can be easily identified by looking for such NPs. In practice, we resort to using automatically produced parse-trees.

We parse the text using the Stanford constituency parser (Chen and Manning, 2014) and look for noun phrases[7] that contain a number but not a noun. This already produces reasonably accurate results, but we found that we can improve further by introducing 10 additional text-based patterns, which were customized based on a development set. These rules look for common cases that are often not captured by the parser. For example, a conjunction pattern involving a number followed by 'or', such as *''eight or nine clubs''*,[8] where 'eight' is an NFH that refers to 'clubs'.

Parsing errors result in false-positives. For example in ''You've had [**one** too many **cosmos**].'', the Stanford parser analyzes 'one' as an NP, despite the head ('cosmos') appearing two tokens later. We cover many such cases by consulting with an additional parser. We use the SPACY dependency parser (Honnibal and Johnson, 2015) and filter out cases where the candidate anchor has a noun as its syntactic head or is connected to its parent via a *nummod* label. We also filter cases where the number is followed or preceded by a currency symbol.

Evaluation We evaluate the rule-based identification on our test set, resulting in 97.4% precision and 93.6% recall. The identification errors are almost exclusively a result of parsing mistakes in the underlying parsers. An example of a false-negative error is in the sentence: ''*The lost **six** belong in Thorn Valley*'', where the dependency parser mistakenly labeled 'belong' as a noun, resulting in a negative classification. An example of a false-positive error is in the sentence: ''*our God is the **one** true God*'' where the dependency parser labeled the head of **one** as 'is'.

[6]One exception are numbers that are part of names ('*Appollo **11**'s your secret weapon?*'), which we do not consider to be NFHs.

[7]Specifically, we consider phrases of type NP, QP, NP-TMP, NX, and SQ.

[8]This phrase can be treated as a gapped coordination construction. For consistency, we treat it and similar cases as NFHs, as discussed in Section 2.2. Another reading is that the entire phrase ''eight or nine'' refers to a single approximate quantity that modifies the noun ''clubs'' as a single unit. This relates to the problem of disambiguating distributive-vs-joint reading of coordination, which we consider to be out of scope for the current work.

	train	dev	test	all
pos	71,821	7865	206	79,884
neg	93,785	10,536	294	104,623
all	165,606	18,401	500	184,507

Table 2: NFH Identification corpus summary. The train and dev splits are noisy and the test set are gold annotations.

4.2 Learning-based Identification

We improve the NFH identification using machine learning. We create a large but noisy data set by considering all the numbers in the corpus and treating the NFHs identified by the rule-based approach as positive (79,678 examples) and all other numbers as negative (104,329 examples). We randomly split the data set into train and development sets in a 90%, 10% split. Table 2 reports the data set size statistics.

We train a linear support vector machine classifier[9] with four features: (1) concatenation of the anchor-span tokens; (2) lower-cased tokens in a 3-token window surrounding the anchor span; (3) part of speech (POS) tags of tokens in a 3-token window surrounding the anchor span; and (4) POS-tag of the syntactic head of the anchor. The features for the classifier require running a POS tagger and a dependency parser. These can be omitted with a small performance loss (see Table 3 for an ablation study on the dev set).

On the manually labeled test set, the full model achieves accuracies of 97.5% precision and 95.6% recall, **surpassing the rule-based approach**.

4.3 NFH Statistics

We use the rule-based positive examples of the data set and report some statistics regarding the NFH phenomenon. The most common anchor of the NFH data set with a very big gap is the token 'one'[10] with 48,788 occurrences (61.0% of the data), while the second most commons is the token 'two' with 6,263 occurrences (8.4%). There is a long tail in terms of the tokens occurrences, with 1,803 unique anchor tokens (2.2% of the NFH data set). Most of the anchors consist of a single token (97.4%), 1.3% contain 2 tokens, and the longest anchor consists of 8 tokens ('Fifteen million sixty one thousand and seventy six.'). The

[9]`sklearn` implementation (Pedregosa et al., 2011) with default parameters.

[10]Lower-cased.

	Precision	Recall	F1
Deterministic (Test)	97.4	93.6	95.5
Full-model (Test)	**97.5**	**95.6**	**96.6**
Full-model (Dev)	**96.8**	**97.5**	**97.1**
- dep	96.7	97.3	97.0
- pos	96.4	97.0	96.7
- dep, pos	95.6	96.1	95.9

Table 3: NFH Identification results.

numbers tend to be written as words (86.7%) and the rest are written as digits (13.3%).

4.4 NFH Identification Data Set

The underlying corpus contains 184,507 examples (2,803,009 tokens), of which 500 examples are gold-labeled and the rest are noisy. In the gold test set, 41.2% of the numbers are NFHs. The estimated quality of the corpus—based on the manual test-set annotation—is 96.6% F1 score. The corpus and the NFH identification models are available at `github.com/yanaiela/num_fh`.

5 NFH Resolution Data Set

Having the ability to identify NFH cases with high accuracy, we turn to the more challenging task of NFH resolution. The first step is creating a gold annotated data set.

5.1 Corpus Candidates

Using the identification methods—which achieve satisfying results—we identify a total of 79,884 NFH cases in the IMDB corpus. We find that a large number of the cases follow a small set of patterns and are easy to resolve deterministically: Four deterministic patterns account for 28% of the NFH cases. The remaining cases are harder. We randomly chose a 10,000-case subset of the harder cases for manual annotation via crowdsourcing. We only annotate cases where the rule-based and learning-based identification methods agree.

Deterministic Cases The four deterministic patterns along with their coverage are detailed in Table 4. The first two are straightforward string matches for the patterns *no one* and *you two*, which we find to almost exclusively resolve to PEOPLE. The other two are dependency-based patterns for partitive (*four [children] of the children*) and copular (*John is the one [John]*) constructions. We collected a total of 22,425 such cases. Although we believe these cases need to be handled by any NFH resolution system, we do not think systems should be evaluated on them. Therefore, we provide these cases as a separate data set.

5.2 Annotation via Crowdsourcing

The FH phenomenon is relatively common and can be understood easily by non-experts, making the task suitable for crowd-sourcing.

The Annotation Task For every NFH anchor, the annotator should decide whether it is a *Reference FH* or an *Implicit FH*. For *Reference*, they should mark the relevant textual span. For *Implicit*, they should specify the implicit head from a closed list. In cases where the missing head belongs to the implicit list, but also appears as a span in the sentence (reference), the annotators are instructed to treat it as a reference. To encourage consistency, we run an initial annotation in which we identified common implicit cases: YEAR (a calendar year, Example (ix) in Table 1), AGE (example x), CURRENCY (Example (xi); although the source of the text suggests US dollars, we do not commit to a specific currency), PERSON/PEOPLE (Example (vi)) and TIME (a daily hour, Example (iii)). The annotators are then instructed to either choose from these five categories; to choose OTHER and provide free-form text; or to choose UNKNOWN in case the intended head cannot be reliably deduced based on the given text.[11] For the *Reference* cases, the annotators can mark any contiguous span in the text. We then simplify their annotations and consider only the syntactic head of their marked span.[12] This could be done automatically in most cases, and was done manually in the few remaining cases. The annotator must choose a single span. In case the answer includes several spans as in examples viii and xiv, we rely on it to surface as a disagreement between the annotators, which we then pass to further resolution by expert annotators.

The Annotation Procedure We collected annotations using Amazon Mechanical Turk (AMT).[13] In every task (HIT in AMT jargon) a sentence

[11] This happens, for example, when the resolution depends on another modality. For example, in our setup using dialogs from movies and TV-series, the speaker could refer to something from the video that isn't explicitly mentioned in the text, such as in ''*Hit the deck, Pig Dog, and give me **37**!*''.

[12] We do provide the entire span annotation as well, to facilitate future work on boundary detection.

[13] To maximize the annotation quality, we restricted the turkers with the following requirements: Complete over 5 K

Pattern	Example	Head	Frequency (%)
no **one**	No **one** cares, dear.	PEOPLE	6.8
you **two**	Are you **two** done with that helium?	PEOPLE	2.3
NUM of *NP*	I had another **one** of those horrible *dreams*!	*dreams*	15.8
X 'be' the **one**	*Theresa* is the **one** who "borrowed" Matt's car.	*Theresa*	2.9

Table 4: Example of NFHs whose heads can be resolved deterministically. The first two patterns are the easiest to resolve. These just have to match as is and their head is the PEOPLE class. The last two patterns depends on a dependency parser and can be resolved by following arcs on the parse tree.

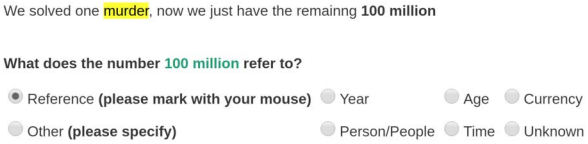

Figure 2: Crowdsourcing task interface on AMT.

with the FH *anchor* was presented (*target sentence*). Each *target sentence* was presented with maximum two dialog turns before and one dialog turn after it. This was the sole context that was shown to avoid exhausting the AMT workers (turkers) with long texts and in the vast majority of the examined examples, the answer appeared in that scope.

Every HIT contained a single NFH example. In cases of more than one NFH per sentence, it was split into 2 different HITs. The annotators were presented with the question: ''What does the number [ANCHOR] refer to?'' where [ANCHOR] was replaced with the actual number span, and annotaters were asked to choose from eight possible answers: REFERENCE, YEAR, AGE, CURRENCY, PERSON/PEOPLE, TIME, OTHER, and UNKNOWN (See Figure 2 for a HIT example). Choosing the REFERENCE category requires marking a span in the text corresponding to the referred element (the missing head). The turkers were instructed to prefer this category over the others if possible. Therefore, in Example (xiv) of Table 1, the *Reference* answers were favored over the PEOPLE answer. Choosing the OTHER category required entering free-form text.

Post-annotation, we unify the *Other* and *Unknown* cases into a single OTHER category.

acceptable HITs, over 95% of their overall HITs being accepted, and completing a qualification for the task.

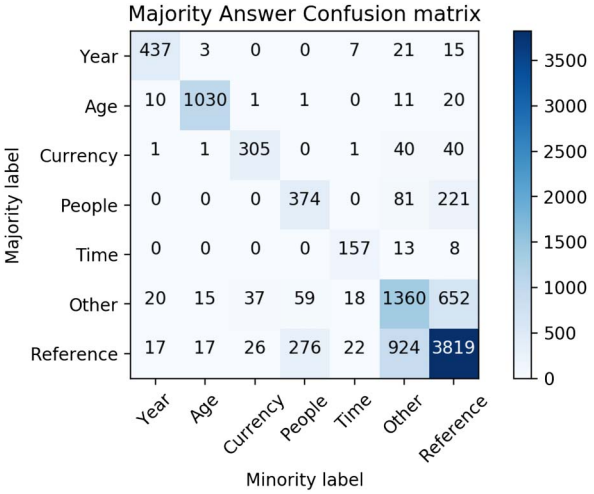

Figure 3: Confusion matrix of the majority annotators on categorical decision.

Each example was labeled by three annotators. On the categorical decision (just the one-of-seven choice, without considering the spans selected for the REFERENCE text and combining the OTHER and UNKNOWN categories), 73.1% of the cases had a perfect agreement (3/3), 25.6% had a majority agreement (2/3), and 1.3% had a complete disagreement. The Fleiss kappa agreement (Fleiss, 1971) is $k = 0.73$, a substantial agreement score. The high agreement score suggests that the annotators tend to agree on the answer for most cases. Figure 3 shows the confusion matrix for the one-of-seven task, excluding the cases of complete disagreement. The more difficult cases involve the REFERENCE class, which is often confused with PEOPLE and OTHER.

5.3 Final Labeling Decisions

Post-annotation, we ignore the free text entry for OTHER and unify OTHER and UNKNOWN into a

single category. However, our data collection process (and the corpus we distribute) contain this information, allowing for more complex task definitions in future work.

The disagreement cases surface genuinely hard cases, such as the ones that follow:

(7) Mexicans have **fifteen**, Jews have thirteen, rich girls have sweet sixteen...

(8) All her *communications* are to Minnesota *numbers*. There's not **one** from California.

(9) And I got to see *Irish*. I think he might be the **one** that got away, or the one that got put-a-way.

The majority of the partial category agreement cases (1,576) are of REFERENCE vs. OTHER/UNKNOWN, which are indeed quite challenging (e.g., Example (9) where two out of three turkers selected the REFERENCE answer and marked *Irish* as the head, and the third turker selected the Person/People label, which is also true, but less meaningful in our perspective).

The final labeling decision was carried out in two phases. First, a categorical labeling was applied using the majority label, while the 115 examples with disagreement (e.g., Example (7), which was tagged as YEAR, REFERENCE ('birthday' which appeared in the context), and OTHER (free text:'special birthday')) were annotated manually by experts.

The second stage dealt with the REFERENCE labels (5,718 cases). We associate each annotated span with the lemma of its syntactic head, and consider answers as equivalent if they share the same lemma string. This results in 5,101 full-agreement cases at the lemma level. The remaining 617 disagreement cases (e.g., Example (8)) were passed to further annotation by the expert annotators. During the manual annotation we allow also for multiple heads for a single anchor (e.g., for viii, xiv in Table 1).

An interesting case in *Reference* FHs is a construction in which the referenced head is not unique. Consider Example (viii) in Table 1: the word 'one' refers to either *men* or *buses*. Another example of such case is Example (xiv) in Table 1 where the word 'two' refers both to *fussy old maid* and to *flashy young man*. Notice that the two cases have different interpretations: The referenced heads in Example (viii) have an *or* relation between them whereas the relation in (xiv) is *and*.

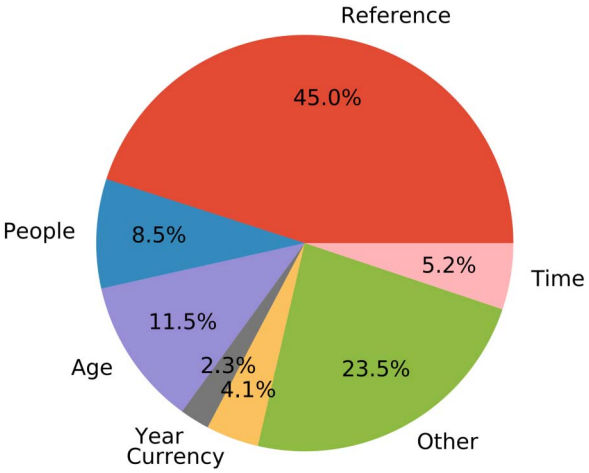

Figure 4: Distribution of NFH types in the NFH Resolution data set.

5.4 NFH Statistics

General We collected a total of 9,412 annotated NFHs. The most common class is REFERENCE (45.0% of the data set). The second common class is OTHER (23.5%), which is the union of original OTHER class, in which turkers had to write the missing head, and the UNKNOWN class, in which no clear answer could be identified in the text. The majority of this joined class is from the UNKNOWN label (68.3%). The rest of the five closed-class categories account for the other 31.5% of the cases. A full breakdown is given in Figure 4. The *anchor* tokens in the data set mainly consist of the token 'one' (49.0% of the data set), with the tokens 'two' and 'three' being the second and third most common. Additionally, 377 (3.9%) of the *anchors* are singletons, which appear only once.

Reference Cases The data set consists of a total of 4,237 REFERENCE cases. The vast majority of them (3,938 cases) were labeled with a single referred element, 238 with two reference-heads, and 16 with three or more.

In most of the cases, the reference span can be found near the anchor span. In 2,019 of the cases, the reference is in the same sentence with the anchor, in 1,747 it appears in a previous/following sentence. Furthermore, in most cases (82.7%), the reference span appears before the anchor and only in 5.1% of the cases does it appear after it. An example of such a case is presented in Example (xiv) in Table 1. In the rest of the cases, references appear both before and after the anchor.

5.5 NFH Resolution Data Set

The final NFH Resolution data set consists of 900,777 tokens containing 9,412 instances of gold-labeled resolved NFHs. The resolution was done by three mechanical turk annotators per task, with a high agreement score ($k = 0.73$).[14] The REFERENCE cases are annotated with at least one referring item. The OTHER class unifies several other categories (None and some other scarce *Implicit* classes), but we maintain the original turker answers to allow future work to apply more fine-grained solutions for these cases.

6 Where's my Head? Resolution Model

We consider the following resolution task: Given a numeric anchor and its surrounding context, we need to assign it a single head. The head can be either a token from the text (for *Reference FH*) or one-of-six categories (the 5 most common categories and OTHER) for *Implicit FH*.[15]

This combines two different kinds of tasks. The REFERENCE case requires selecting the most adequate token over the text, suggesting a similar formulation to coreference resolution (Ng, 2010; Lee et al., 2018) and implicit arguments identification (Gerber and Chai, 2012; Moor et al., 2013). The implicit case requires selection from a closed list, a similar formulation to word-tagging-in-context tasks, where the word (in our case, span) to be tagged is the anchor. A further complication is the need to weigh the different decisions (*Implicit* vs. *Reference*) against each other. Our solution is closely modeled after the state-of-the-art coreference resolution system of Lee et al. (2017).[16] However, the coreference-centric architecture had to be adapted to the particularities of the NFH task. Specifically, (a) the NFH resolution does not involve cluster assignments, and (b) it requires handling the *Implicit* cases in addition to the *Reference* ones.

The proposed model combines both decisions, a combination that resembles the copy-mechanisms in neural MT (Gu et al., 2016) and the Pointer Sentinel Mixture Model in neural LM (Merity et al., 2016). As we only consider referring mentions as single tokens, we discarded the original models' features that handled the multi-span representation (e.g., the Attention mechanism). Furthermore, as the Resolution task already receives a numeric anchor, it is redundant to calculate a mention score. In preliminary experiments we did try to add an antecedent score, with no resulting improvement. Our major adaptations to the Lee et al. (2017) model, described subsequently, are the removal of the redundant components and the addition of an embedding matrix for representing the *Implicit* classes.

6.1 Architecture

Given an anchor, our model assigns a score to each possible anchor–head pair and picks the one with the highest score. The head can be either a token from the text (for the *Reference* case) or one-of-six category labels (for the *Implicit* case). We represent the anchor, each of the text tokens and each category label as vectors.

Each of the implicit classes $c_1, \ldots, c_6$ is represented as an embedding vector $\mathbf{c_i}$, which is randomly initialized and trained with the system.

To represent the sentence tokens ($\mathbf{t_i}$), we first represent each token as a concatenation of the token embedding and the last state of a character long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997):

$$\boldsymbol{x_i} = [\mathbf{e_i}; LSTM(\mathbf{e_{i_{c1:ct}}})]$$

where $\mathbf{e_i}$ is the ith token embedding and $\mathbf{e_{i_{c_j}}}$ is the jth character of the ith token. These representations are then fed into a text-level biLSTM resulting in the contextualized token representations $\mathbf{t_i}$:

$$\mathbf{t_i} = BILSTM(\mathbf{x_{1:n}}, i)$$

Finally, the *anchor*, which may span several tokens, is represented as the average over its contextualized tokens.

$$\boldsymbol{a} = \frac{1}{j-i+1}\sum_{k=i}^{j} \boldsymbol{t_k}$$

[14] The *Reference* cases were treated as a single class for computing the agreement score.

[15] This is a somewhat simplified version of the full task defined in Section 3. In particular, we do not require specification of the head in case of OTHER, and we require a single head rather than a list of heads. Nonetheless, we find this variant to be both useful and challenging in practice. For the few multiple-head cases, we consider each of the items in the gold list to be correct, and defer a fuller treatment for future work.

[16] Newer systems such as Lee et al. (2018) and Zhang et al. (2018) show improvements on the coreference task, but use components that focus on the clustering aspect of coreference, which are irrelevant for the NFH task.

We predict a score $s(h, a)$ for every possible head-anchor pair, where $h \in \{c_1, \ldots, c_6, t_1, \ldots, t_n\}$ and $\mathbf{h_i}$ is the corresponding vector. The pair is represented as a concatenation of the head, the anchor and their element-wise multiplication, and scored with a multi-layer perceptron:

$$s(h, a) = MLP([\mathbf{h}; \mathbf{a}; \mathbf{h} \odot \mathbf{a}])$$

We normalize all of the scores using softmax, and train to minimize the cross-entropy loss.

Pre-trained LM To take advantage of the recent success in pre-trained language models (Peters et al., 2018; Devlin et al., 2018) we also make use of ELMo contextualized embeddings instead of the embedding matrix and the character LSTM concatentation.

6.2 Training Details

The character embedding size is 30 and their LSTM dimension is 10. We use Google's pre-trained 300-dimension w2v embeddings (Mikolov et al., 2013) and fix the embeddings so they don't change during training. The text-level LSTM dimension is 50. The *Implicit* embedding size is the same as the BiLSTM output, 100 units. The MLP has a single hidden layer of size 150 and uses $tanh$ as the non-linear function. We use dropout of 0.2 on all hidden layers, internal representation, and tokens representation. We train using the Adam optimizer (Kingma and Ba, 2015) and a learning rate of 0.001 with early stopping, based on the development set. We shuffle the training data before every epoch. The annotation allows more than one referent answer per anchor; in such case, we take the closest one to the anchor as the answer for training, and allow either one when evaluating. The experiments using ELMo replaced the pre-trained word embeddings and character LSTM. It uses the default parameters in the AllenNLP framework (Gardner et al., 2017), with 0.5 dropout on the network, without gradients update on the contextualized representation.

6.3 Experiments and Results

Data Set Splits We split the data set into train/development/test, containing 7,447, 1,000, and 1,000 examples, respectively. There is no overlap of movies/TV-shows between the different splits.

Model	Reference	Implicit
Oracle (Reference)	70.4	-
+ Elmo	81.2	-
Oracle (Implicit)	-	82.8
+ Elmo	-	90.6
Model (full)	61.4	69.2
+ Elmo	73.0	80.7

Table 5: NFH Resolution accuracies for the *Reference* and *Implicit* cases on the development set. **Oracle (Reference)** and **Oracle (Implicit)** assume an oracle for the implicit vs. reference decisions. **Model (full)** is our final model.

Metrics We measure the model performance of the NFH head detection using accuracy. For every example, we measure whether the model successfully predicted the correct label or not. We report two additional measurements: Binary classification accuracy between the *Reference* and *Implicit* cases and a multiclass classification accuracy score, which measures the class-identification accuracy while treating all REFERENCE selections as a single decision, regardless of the chosen token.

Results We find that 91.8% of the *Reference* cases are nouns. To provide a simple baseline for the task, we report accuracies solely on the *Reference* examples (ignoring the *Implicit* ones) when choosing one of the surrounding nouns. Choosing the first noun in the text, the last one or the closest one to the anchor leads to scores of 19.1%, 20.3%, and 39.2%.

We conduct two more experiments to test our model on the different FH kinds: *Reference* and *Implicit*. In these experiments we assume an oracle that tells us the head type (*Implicit* or *Reference*) and restricts the candidate set for the correct kind during both training and testing. Table 5 summarizes the results for the oracle experiments as well as for the full model.

The final models accuracies are summarized in Table 6. The complete model trained on the entire training data achieves 65.6% accuracy on the development set and 60.8% accuracy on the test set. The model with ELMo embeddings (Peters et al., 2018) adds a significant boost in performance and achieves 77.2% and 74.0% accuracy on the development and test sets, respectively.

The development-set binary separation with ELMo embeddings is 86.1% accuracy and categorical separation is 81.9%. This substantially outperforms all baselines, but still lags behind

Model	Development	Test
Base	65.6	60.8
+ Elmo	**77.2**	**74.0**

Table 6: NFH Resolution accuracies on the development and test sets.

the oracle experiments (*Reference*-only and *Implicit*-only).

As the oracle experiments perform better on the individual *Reference* and *Implicit* classes, we experimented with adding an additional objective to the model that tries to predict the oracle decision (implicit vs. reference). This objective was realized as an additional loss term. However, this experiment did not yield any performance improvement.

We also experimented with linear models, with features based on previous work that dealt with antecedent determination (Ng et al., 2005; Liu et al., 2016) such as POS tags and dependency labels of the candidate head, whether the head is the closest noun to the anchor, and so forth. We also added some specific features that dealt with the Implicit category, for example binarization of the anchor based on its magnitude (e.g., < 1, < 10, < 1600, < 2100), if there was another currency mention in the text, and so on. None of these attempts surpassed the 28% accuracy on the development set. For more details on these experiments, see Appendix A.

6.4 Analysis

The base model's results are relatively low, but gain a substantial improvement by adding contextualized embeddings. We perform an error analysis on the ELMo version, which highlights the challenges of the task.

Figure 5 shows the confusion matrix of our model and Table 7 lists some errors from the development set.

Pattern-Resolvable Error Cases The first three examples in Table 7 demonstrate error cases that can be solved based on text-internal cues and ''complex-pattern-matching'' techniques. These can likely be improved with a larger training set or improved neural models.

The errors in rows 1 and 2 might have caused by a multi-sentence patterns. A possible reason for the errors is the lack of that pattern in the training data. Another explanation could be a magnitude

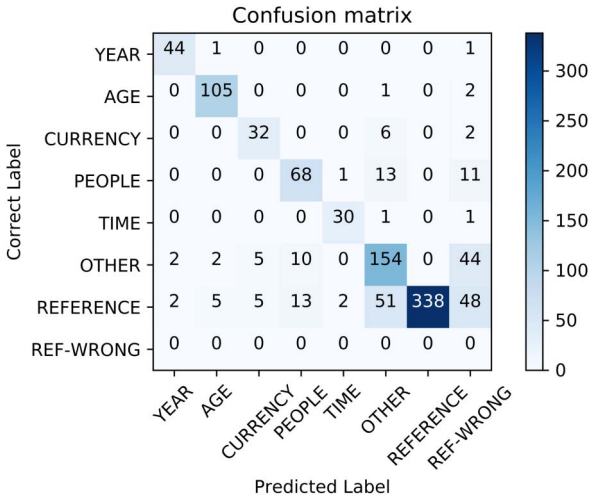

Figure 5: Confusion matrix of the model. Each row/column corresponds to a gold/predicted label respectively. The last one (REF-WRONG), is used for indicating an erroneous choice of a *Reference* head.

bias, where in row 1, **One** in the beginning of a sentence usually refer to PEOPLE, whereas in row 2, **Five** is more likely to refer to an AGE.

In row 3, the model has to consider several cues from the text, such as the phrase ''*a hundred dollars*'' which contains the actual head and is of a similar magnitude to the anchor. In addition, the phrase: ''*it was more around*'' gives a strong hint on a previous reference.

Inference/Common Sense Errors Another category of errors includes those that are less likely to be resolved with pattern-based techniques and more data. These require common sense and/or more sophisticated inferences to get right, and will likely require a more sophisticated family of models to solve.

In row 4, **one** refers to *dad*, but the model chose *sisters*. These are the only nouns in this example, and, with the lack of any obvious pattern, a model needs to understand the semantics of the text to identify the missing head correctly.

Row 5 also requires understanding the semantics of the text, and some understanding of its discourse dynamic; where a conversation between the two speakers takes place, with a reply of *Krank* to *L'oncle Irvin*, that the model missed.

In Row 6, the model has difficulty collecting the cues in the text that refer to an unmentioned person, and therefore the answer is PEOPLE, but the model predicts OTHER.

Finally, in Row 7 we observe an interesting case of overfitting, which is likely to originate

	Text	Predicted	Truth
1	**Dreadwing:** This will be my gift to the Dragon Flyz, my farewell *gift* . **One** that will keep giving and giving and giving.	PEOPLE	*gift*
2	**David Rossi:** How long? **Harrison Scott:** A year . Maybe **five**. It's hard to keep track without a watch.	AGE	YEAR
3	**Henry Fitzroy:** a hundred *dollars*, that's all it takes for you to risk your life ? **Vicki Nelson:** Actually, it was more around **98**...	OTHER	*dollar*
4	**Evelyn Pons:** He might be my legal *dad* , too! **Paula Novoa Pazos:** No, because we're not sisters , but you can look for another one. **Evelyn Pons:** How did you look for **one**?	*sisters*	*dad*
5	**L'oncle Irvin:** A *soul* . **Krank:** Because you believe you have **one**? You don't even have a body .	*body*	*soul*
6	**Jenny:** Head in the clouds, that **one**. I don't know why you're so sweet on him.	OTHER	PEOPLE
7	**Officer Mike Laskey:** I can't do that. **Joss Carter:** Do you really wanna test me? 'Cause I've got a shiny new **1911** [...]	YEAR	OTHER

Table 7: Erroneous example predictions from the development data. Each row represents an example from the data. The redder the words, the higher their scores. The two last columns contain the model prediction and the gold label. Uppercase means the label is from the IMPLICIT classes, otherwise it is a REFERENCE in lowercase.

from the word-character encoding. As the anchor - **1991** is a four-digit number, which are usually used to describe YEARS, its representation receives a strong signal for this label, even though the few words which precede it (*a shiny new*) are not likely to describe a YEAR label.

7 Related Work

The FH problem has not been directly studied in the NLP literature. However, several works have dealt with overlapping components of this problem.

Sense Anaphora The first, and most related, is the line of work by Gardiner (2003), Ng et al. (2005), and Recasens et al. (2016), which dealt with sense anaphoric pronouns (''Am I a *suspect*? - you act like **one**'', cf. Example (4)). *Sense anaphora*, sometimes also referred to as *identity of sense anaphora*, are expressions that inherit the sense from their antecedent but do not denote the same referent (as opposed to coreference). The sense anaphora phenomena also cover numerals, and significantly overlap with many of our NFH cases. However, they do not cover the *Implicit* NFH cases, and also do not cover cases where the target is part of a co-referring expression (''I met *Alice and Bob*. The **two** seem to get along well.'').

In terms of computational modeling, the sense anaphora task is traditionally split into two subtasks: (i) identifying anaphoric targets and disambiguating their sense; and (ii) resolving the target to an antecedent. Gardiner (2003) and Ng et al. (2005) perform both tasks, but restrict themselves to *one anaphora* cases and their noun-phrase antecedents. Recasens et al. (2016), on the other hand, addressed a wider variety of sense anaphors (e.g., *one*, *all*, *another*, *few*, *most*—a total of 15 different senses, including *numerals*). Recasens et al. (2016) annotated a corpus of a third of the English OntoNotes (Weischedel et al., 2011) with sense anaphoric pronouns and their antecedents. Based on this data set, they introduce a system for distinguishing anaphoric from non-anaphoric usages. However, they do not attempt to resolve any target to its antecedent. The non-anaphoric examples in their work combines both our *Implicit* class, as well as other non-anaphoric examples indistinguishably, and therefore are not relevant for our work.

In the current work, we restrict ourselves to numbers and so cover only part of the sense-anaphora cases handled in Recasens et al. (2016). However, in the categories we do cover, we do not limit ourselves to anaphoric cases (e.g., Examples (3), (4)) but include also non-anaphoric cases that occur in FH constructions (e.g., Examples (1), (2)) and are interesting on their own right. Furthermore, our models not only identify the anaphoric cases but also attempt to resolve them to their antecedent.

Zero Reference In *zero reference*, the argument of a predicate is missing, but it can be easily understood from context (Hangyo et al., 2013). For example, in the sentence: *''There are two roads to eternity, a straight and narrow __ , and a broad and crooked __ ''* have a zero-anaphoric relationship to ''two roads to eternity'' (Iida et al., 2006). This phenomenon is usually discussed as the context of *zero pronouns*, where a pronoun is what is missing. It occurs mainly in pro-drop languages such as Japanese, Chinese, and Italian, but has also been observed in English, mainly in conversational interactions (Oh, 2005). Some, but not all, zero-anaphora cases result in FH or NFH instances. Similarly to FH, the omitted element can appear in the text, similar to our *Reference* definition (zero **endophora**), or outside of it, similar to our *Implicit* definition (zero **exophora**). Identification and resolution of this has attracted considerable interest mainly in Japanese (Nomoto and Nitta, 1993; Hangyo et al., 2013; Iida et al., 2016) and Chinese (Chen and Ng, 2016; Yin et al., 2018a,b), but also in other languages (Ferrández and Peral, 2000; Yeh and Chen, 2001; Han, 2004; Kong and Zhou, 2010; Mihăilă et al., 2010; Kopeć, 2014). However, most of these works considered only the zero endophora phenomenon in their studies, and even those who did consider zero exophora (Hangyo et al., 2013), only considered the author/reader mentions, for example, *''liking pasta (ϕ) eats (ϕ) every day''* (translated from Japanese). In this study, we consider a wider set of possibilities. Furthermore, to the best of our knowledge, we are the first to tackle (a subset-of) zero anaphora in English.

Coreference The coreference task is to find within a document (or multiple documents) all the corefering spans that form cluster(s) of the same mention (which are the anaphoric cases as described above). The FHs resolution task, apart from the non-anaphoric cases, is to find the correct anaphora reference of the target span. The span identification component of our task overlaps with the coreference one (see Ng [2010] for a thorough summary on the NP coreference resolution and Sukthanker et al. [2018] for a comparison between coreference and anaphora). Although the span search resemblance, the key conceptual distinctions is that FHs allow the anaphoric span to be non co-referring.

Recent work on coreference resolution (Lee et al., 2017) propose an end-to-end neural architecture that results in a state-of-the-art performance. The work of Peters et al. (2018), Lee et al. (2018), and Zhang et al. (2018) further improve on their the scores with pre-training, refining span representation and using biaffine attention model for mention detection and clustering. Although these models cannot be applied to the NFH task directly, we propose a solution based on the model of Lee et al. (2017), which we adapt to incorporate the implicit cases.

Ellipsis The most studied type of ellipsis is the Verb Phrase Ellipsis (VPE). Although the following refers to this line of studies, the task and resemblance to the NFH task hold up to the other types of ellipsis as well (gapping [Lakoff and Ross, 1970], sluicing [John, 1969], nominal ellipsis [Lobeck, 1995], etc.). VPE is the anaphoric process where a verbal constituent is partially or totally unexpressed but can be resolved through an antecedent from context (Liu et al., 2016). For example, in the sentence: ''His wife also *works for the paper*, as **did** his father'', the verb **did** is used to represent the verb phrase *works for the paper*. The VPE resolution task is to detect the target word which creates the ellipsis and the anaphoric verb phrase which it depicts. Recent work (Liu et al., 2016; Kenyon-Dean et al., 2016) tackles this problem by dividing it into two main parts: Target detection and antecedent identification.

Semantic Graph Representations Several semantic graph representation cover some of the cases we consider. Abstract Meaning Representation is a graph-based semantic representation for language (Pareja-Lora et al., 2013). It covers a wide range of concepts and relations. Five of those concepts: *Year*, *age*, *monetary-quantity*, *time*, and *person* correlate to our implicit classes: YEAR, AGE, CURRENCY, TIME, and PEOPLE, respectively.

The UCCA semantic representation (Abend and Rappoport, 2013) explicitly marks missing information, including the REFERENCE NFH cases, but not the IMPLICIT ones.

8 Conclusions

Empty elements are pervasive in text, yet do not receive much research attention. In this work, we tackle a common phenomenon that did not receive previous treatment. We introduce the FH

identification and resolution tasks and focus on a common and important FH subtype: The NFH. We demonstrate that the NFH is a common phenomenon, covering over 40% of the number appearances in a large dialog-based corpus and a substantial amount in other corpora as well ($> 20\%$). We create data sets for the NFH identification and resolution tasks. We provide an accurate method for identifying the NFH constructions and a neural baseline for the resolution task. The resolution task proves challenging, requiring further research. We make the code and data sets available to facilitate such research (`github.com/yanaiela/num_fh`).

Acknowledgments

We would like to thank Reut Tsarfaty and the Bar-Ilan University NLP lab for the fruitful conversation and helpful comments. The work was supported by the Israeli Science Foundation (grant 1555/15) and the German Research Foundation via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238. Sofia.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012, May. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268. Trento.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 778–788.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

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

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 166–172.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Mary Gardiner. 2003. Identifying and resolving oneanaphora. Unpublished Honours thesis, Macquarie University, November.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. arXiv preprint `arXiv:1803.07640`.

Matthew Gerber and Joyce Y. Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Na-Rae Han. 2004. Korean null pronouns: Classification and annotation. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 33–40.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese zero reference resolution considering exophora and

author/reader mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 924–934.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon.

Rodney Huddleston and Geoffrey K. Pullum. 2002. The Cambridge Grammar of English. *Language*. Cambridge: Cambridge University Press, pages 1–23.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1244–1254.

Ross John. 1969. Guess who. In *Proceedings of the 5th Chicago Linguistic Society*, pages 252–286.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2016. Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. In *Proceedings of EMNLP*, pages 1734–1743.

Diederik P. Kingma and Lei Ba. 2015. J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891.

Mateusz Kopeć. 2014. Zero subject detection for Polish. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 221–225.

George Lakoff and John Robert Ross. 1970. Gapping and the order of constituents. *Progress in Linguistics: A Collection of Papers*, 43:249.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv: 1707.07045*.

Kenton Lee, Luheng He, and Luke S. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Iddo Lev, Bill MacCartney, Christopher D Manning, and Roger Levy. 2004. Solving logic puzzles: From robust processing to precise semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, pages 9–16.

Zhengzhong Liu, Edgar Gonzàlez Pellicer, and Daniel Gillick. 2016. Exploring the steps of verb phrase ellipsis. In *CORBON@ HLT-NAACL*, pages 32–40.

Anne C. Lobeck. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*, Oxford University Press on Demand.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv: 1609.07843*.

Claudiu Mihăilă, Iustina Ilisei, and Diana Inkpen. 2010. To be or not to be a zero pronoun: A machine learning approach for romanian. *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, 303–316.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tatjana Moor, Michael Roth, and Anette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 369–375, Potsdam.

Hwee Tou Ng, Yu Zhou, Robert Dale, and Mary Gardiner. 2005. A machine learning approach to identification and resolution of one-anaphora. In *International Joint Conference on Artificial Intelligence*, volume 19, page 1105.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411.

Tadashi Nomoto and Yoshihiko Nitta. 1993. Resolving zero anaphora in japanese. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*, pages 315–321.

Sun-Young Oh. 2005. English zero anaphora as an interactional resource. *Research on Language and Social Interaction*, 38(3):267–302.

Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper. 2013. Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Marta Recasens, Zhichao Hu, and Olivia Rhinehart. 2016. Sense anaphoric pronouns: Am i one? In *CORBON@ HLT-NAACL*, pages 1–6.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Georgios P. Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. *arXiv preprint arXiv:1805.08154*.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and coreference resolution: A review. *arXiv preprint arXiv:1805.11824*.

Andrew Trask, Felix Hill, Scott Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. *arXiv preprint arXiv:1808.00508*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, and Robert Belvin. 2011. Ontonotes release 4.0. LDC2011T03, Philadelphia, PA: Linguistic Data Consortium.

Ching-Long Yeh and Yi-Jun Chen. 2001. An empirical study of zero anaphora resolution in chinese based on centering model. In *Proceedings of Research on Computational Linguistics Conference XIV*, pages 237–251.

Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2018a. Deep reinforcement learning for Chinese zero pronoun resolution. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 569–578.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018b. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23.

Rui Zhang, Cicero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. *arXiv preprint arXiv:1805.04893*.

A Details of Linear Baseline Implementation

This section lists the features used for the linear baseline mentioned in Section 6.3. The features are presented in Table 8. We used four type of features: (1) **Label** features, making use of parsing labels of dependency and POS-taggers, as well as simple lexical features of the *anchor*'s window. (2) **Structure** features, incorporating structural information from the sentence and the *anchor*'s spans. (3) **Match** features test for specific patterns in the text, and (4) **Other**, not-categorized features.

Type	Feature Description
Labels	Anchor & head lemma
	2 sized window lemmas
	2 sized window POS tags
	Dependency edge of target
	Head POS tag
	Head lemma
	Left most child lemma of anchor head
	Children of syntactic head
Structure	Question mark before or after the anchor
	Sentence length bin ($< 5 < 10 <$)
	Span length bin (1, 2 or more)
	Hyphen in anchor span
	Slash in anchor span
	Apostrophe before or after the span
	Apostrophe + 's' after span
	Anchor is ending the sentence
Match	Whether the text contains a currency expression
	Whether the text contains a time expression
	Entity exists in the sentence before the target
Other	Target size bin ($< 1 < 10 < 100 < 1600 < 2100 <$)
	The number shape (digit or written text)

Table 8: Features used for linear classifier.

We used the features described above to train a linear support vector machine classifier on the same splits.