ReferenceNet: a semantic-pragmatic network for capturing reference relations

Piek Vossen, Marten Postma, Filip Ilievski
VU University Amsterdam, Netherlands
{piek.vossen, m.c.postma, f.ilievski}@vu.nl

Abstract

In this paper, we present ReferenceNet: a semantic-pragmatic network of reference relations between synsets. Synonyms are assumed to be exchangeable in similar contexts and also word embeddings are based on sharing of local contexts represented as vectors. Co-referring words, however, tend to occur in the same topical context but in different local contexts. In addition, they may express different concepts related through topical coherence, and through author framing and perspective. In this paper, we describe how reference relations can be added to WordNet and how they can be acquired. We evaluate two methods of extracting event coreference relations using WordNet relations against a manual annotation of 38 documents within the same topical domain of gun violence. We conclude that precision is reasonable but recall is lower because the WordNet hierarchy does not sufficiently capture the required coherence and perspective relations.

1 Introduction

Synonyms from the same synset (Fellbaum, 1998) are assumed to be exchangeable in contexts. Similarly, word embeddings are based on sharing of contexts represented as vectors (Mikolov et al., 2013; Baroni et al., 2014). Both synsets and word embeddings capture some variation in language, but they do not fully capture variation in reference and coreference. Reference relations are different in that they cross local (sentence) contexts. We typically tell stories in discourse in which entities or events play different roles and reflect different phases in relation to the same incident (the topic of the story). Furthermore, authors may frame these entities and events differently either within the same story or across different stories. We can thus consider a story as a larger topical context within which co-referring expressions occur in different local contexts. Each local context of a co-referring expression may represent a different concept. The set of local contexts within a topical context is therefore expected to express not only similarity, but also topical coherence and author framing and perspective.

The next two examples show two fragments from two news articles that make reference to the same incident (topical coherence) in which a man shot several people in a bar in Pittsburgh. The first fragment is published shortly after the incident when the suspect has not yet been identified. The second fragment is published later after the suspect was identified, found guilty and sentenced to prison (changing perspective).

> Investigators continue to look for suspects after one person was killed and four others were injured when gunfire erupted overnight at a bar in Homewood . ..... Several witnesses , [...] They believe the gunman was not searched by the four security guards who left the business before police arrived .

> Man Gets 15 - 30 Years For Deadly Shooting At Homewood Bar . PITTSBURGH (KDKA) A man has pleaded guilty in a 2014 shooting that left four men injured and one dead . Cornell Poindexter , 30 , appeared in court Monday and pleaded guilty to one count of 3rd degree murder , four counts of aggravated assault and one count of person not to possess a firearm . ... According to our partners at The Pittsburgh Post - Gazette , 23-year - old Corey Clark was originally accused of being the gunman , but those charges were dropped .

The following words and expressions are used to make reference to the incident or parts of the incident: *killed*, *injured*, *gunfire erupted* (first fragment) and *deadly shooting*, *shooting*, *left injured and dead*, *murder*, *aggravated assault* (second fragment). The references to the shooter are made through *suspects*, *gunman* and through *man*, *Cornell Poindexter*, *person not to possess a firearm*, *23-year - old Corey Clark* and *gunman* respectively. References to the events differ across the text due to the legal view, e.g. *murder*, whereas the entity references differ due to having more knowledge on the identity of the suspects and the fact that one suspect turned out to be innocent and another was convicted. Making reference is more than similarity of meaning, as it is also governed by pragmatic principles related to information sharing, relevance, salience, and framing. In the different sentences of a discourse, we tend to tell different things about the same referents. These sentences thus represent different local contexts, which are connected through the topical context of the story that is told. From a language understanding and generation

perspective, WordNet synsets and word embeddings are not expected to provide sufficient information to predict usage of one expression over the other, or to infer from the referential usage of expressions what is the semantic implicature (coherence or framing).

We therefore propose to add a layer to WordNet, that captures variation in making reference within a topical context across different synsets or word embeddings that represent local contexts. In this paper, we describe how these relations can be acquired as a **ReferenceNet**. The relations in a ReferenceNet exceed the notion of synonymy and partially also hyponymy and capture a broader range of roles, perspectives, and also different phases of processes. Referential relations can not only help detecting coreference and coherence relations, but also help distinguishing roles from rigid types which is important for further ontologisation of semantic networks, and capturing different ways of framing the same things.

This paper is structured as follows. In section 2, we discuss related work and present the motivation for adding referential relations to WordNet. In section 3, we define the model for expressing these relations. We present two approaches to acquire these relations in sections 4 and 5. Section 6 describes the evaluation data created and section 7 contains the evaluation results. Finally, we conclude and discuss future work in section 8.

2 Related work and motivation

Reference and identity have been discussed extensively in the philosophical literature (Quine and Van, 1960; Kripke, 1972; Putnam, 1973; Frege, 1892; Rast and others, 2007; Wittgenstein, 2010). The linguistic field of lexical pragmatics (Levinson, 1983; Matsumoto, 1995; Blutner, 1998; Weigand, 1998) tries to explain variation in reference as a function of pragmatic principles such as the Gricean maxims (Grice et al., 1975): be maximally informative but no more informative than necessary. Variation of form is partly explained through pragmatic licensing: the least complex form that yields the most salient implicatures is preferred among all forms that can potentially yield these implications. Such principles may predict how we make reference to real-world situations using certain words and expressions and not others, given the shared knowledge we have about these situations.

The way we make reference is however not only determined by efficiency, salience and information sharing, but also by the framing of referents by the author. FrameNet (Baker et al., 2003) is a large resource that describes different ways in which situations can be framed. Frames and frame elements in FrameNet are very specific and typically model the specific realisation of lexical units in texts. It is not clear how to generalise over the specific frames (what do they share or have in common) nor to derive from the database which combinations of frames can be expected within specific topical contexts.

We believe it is worthwhile to investigate empirically the actual referential relations that occur within topical contexts at a large scale, as well as to describe the observed lexical variation according to both pragmatic principles of quality and efficiency and framing principles. We therefore propose a ReferenceNet as a data structure that captures the observed coherence and framing relations between WordNet synsets. ReferenceNet therefore extends WordNet with a new orthogonal relation, which is less strict and limited than FrameNet, and more specific than for instance WordNet Domains (Strapparava et al., 2004). We argue that such data can be potentially very valuable, as it enables our community: 1. to investigate the semantic-pragmatic implications of making reference 2. to learn about the contextual roles and perspectives that govern the usage of these words and expressions, and 3. to improve the detection of these relations by coreference systems.

3 The ReferenceNet model

We define a ReferenceNet as a collection of **ReferenceSets**. A ReferenceSet consists of:

1. the **words and expressions** that have been used to make reference to the same individual in a topical context
2. the list of different **synsets** associated with these words and expressions in this context
3. the **type of topical context** in which the reference relation was observed

As synsets represent concepts, the variety of synsets reflects the range of things or denotation that is captured in a single ReferenceSet. As this range is not ontologically defined, it reflects the *typical* ways in which we frame and conceptualize individuals in topical situations. Typically, these synsets cannot be disjoint (mutually exclusive): they should either belong to the same hypernym chain (being more or less specific), or should be orthogonal according to formal ontological criteria (Guarino, 1999). A ReferenceSet may consist of one or more synsets and the same synset may participate in more than one reference set, thus constituting a 'many-to-many' relation. In addition to the synset of the expression, we also need to record the actual form or synonym from the synset that was used to make reference.[1] As the constraints for making reference with different expressions and different concepts are mildly ontological, it make sense to register the referential usage of expressions and synsets using counters.

[1]Note that in case of proper names, we abstract from the proper name to the most specific WordNet synset or entity type of which the entity is an instance. When building ReferenceSets from large text collections it makes sense to leave out the proper name references, as we would otherwise include all people's names in the ReferenceNet.

Finally, ReferenceSets include an attribute to mark the type of topical context within which referential variation is observed. The topical context underlies the coherence relations within a discourse. Moreover, it explains the variation in making reference to the same entities and events either through shifting roles, phases, and aspects, or through framing by the author. The topical context allows us to abstract from references to individual entities and events, by generalising the observations to the surface forms and synsets. For example, the same person may be referenced during *school*, *family*, *leisure*, or at *work*. It makes little sense to combine all the references to the same person in a single ReferenceSet. Instead, we gather reference to individuals across all different incidents within the same type of topical context. This captures our general ways of framing persons and events within these topics and according to some topical schema. ReferenceSets thus will reflect which synonyms from which synsets are used how frequently to make reference within the same topical context.

Figure 1 shows two examples of a ReferenceSet for the two texts in the introduction that report on the same incident and thus the same topical context of *gun-violence*. We see separate ReferenceSets for the *shooter* and for the *shooting*. Each ReferenceSet consists of a list of *synset-ref* elements.[2] The *synset-ref* element has attributes for the CILI identifier *iid* (Bond et al., 2016; Vossen et al., 2016b), the language specific WordNet synset, and the *corefcount* attribute to express how often this entity was mentioned in the text. Each *synset-ref* contains a list of *surface-form* elements with the surface form and its observed token frequency of making reference.

We can see that the words span different synsets and also different parts-of-speech tags. The first ReferenceSet exhibits the perspective of the *shooter* and the *suspect* before the trial. We abstracted from the actual names of the people through a separate element and counter *proper-name*. The second ReferenceSet shows different granularities of the event: the overall *incident*, the *shooting*, *hitting* and the *outcome*, and it shows the legal judgments: *murder*, *assault*. This illustrates that the reference relations are often orthogonal to hypernym relations.

ReferenceSets as in Figure 1 can be derived from collections of texts in which coreference relations are resolved across documents making reference to the same incident, involving the same entities and events. ReferenceSet can then be formed by aggregation across incidents of the same topic type, based on sufficient overlap between surface forms and synsets of incidents. We discuss methodologies for building a ReferenceNet in detail in the next section.

[2] At the end of each *synset-ref element* we list the corresponding WordNet synonyms as a comment.

4 Methodologies for building a ReferenceNet

Semantic parsing aims at generating a representation of entities and events from their mentions throughout this text. It relies on a broad range of NLP techniques such as tokenization, parsing, named-entity recognition and linking, and semantic role labeling. Coreference modules often operate on top of the output of such modules. Words and phrases that make reference to the same individual or event are coreferential. If different documents report on the same entities, these would ideally result in cross-document coreference. Applying coreference modules to large collections of texts potentially gives us the different ways in which people make reference to the same entities and events in the world. If for all these referential expressions, we would also know the WordNet synsets, we can abstract from coreferential mentions to their synsets and derive ReferenceSets for the semantic *types* of referents. This requires Word Sense Disambiguation (WSD) to run in addition to establishing coreference relations.

The feasibility of this approach depends on the quality of all the underlying modules (among which WSD) as well as the quality of the coreference modules. A distinction can be made between nominal/entity and event coreference, as they are defined and approached differently by different research groups. As we are primarily interested in the topical coherence underlying texts in this paper, we focus in this paper on event coreference and leave nominal or entity coreference for future work. We discuss two methods for obtaining event ReferenceNet data from text collections using semantic parsing: 1) text-to-data and 2) data-to-text.

Text-to-data involves semantic text parsing without knowing the referents a priori and without knowing which texts report on the same incident. It therefore relies on high-quality cross-document event coreference resolution and it is computationally very expensive as it requires comparing all event mentions with each other (within and across documents). Automatic event coreference is a difficult task (Hovy et al., 2013) and made little progress over the years. To compare different approaches on the ECB+ dataset (Cybulska and Vossen, 2014), Yang et al. (2015) reimplemented state-of-the-art algorithms proposed by Bejan and Harabagiu (2010) and Chen and Ji (2009), as well as their own approach. They report 58.7 CoNLL-F1 (Luo et al., 2014) on ECB+ for their own approach, compared to 53.6 CoNLL-F1 for (Bejan and Harabagiu, 2010) and 55.2 CoNLL-F1 for (Chen and Ji, 2009). They obtained their results however only after boosting event detection from an original 65F to 95F by training a separate event detection system on part of the ECB+ data. Without such nearly perfect event detection, their results are much lower. All three approaches are clustering approaches over the dataset using event mentions as input. Likewise, they can only recover coreference relations between mentions that match local structural

```
<ReferenceSet topic="gun-violence">
  <synset-ref corefcount="2" wid="pwn30:eng-10287213-n" iid="i90357"> <!-- gunman, gun -->
    <surface-form "tokencount="2">gunman</surface-form>
  </synset-ref>
  <synset-ref corefcount="2" wid="pwn30:eng-10152083-n" iid="i91182"> <!-- man, adult_male -->
    <surface-form "tokencount="2">man</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-10681383-n" iid=" i93471"> <!-- suspect -->
    <surface-form "tokencount="1">suspect</surface-form>
  </synset-ref>
  <synset-ref corefcount="3" wid="pwn30:eng-00007846-n" iid=" i35562"> <!-- person, individual, someone, somebody, mortal, soul -->
    <surface-form "tokencount="1">person</surface-form>
    <proper-name "tokencount="2"/>
  </synset-ref>
</ReferenceSet>

<ReferenceSet topic="gun-violence">
  <synset-ref corefcount="1" wid="pwn30:eng-00355365-v" iid="i23513"> <!-- kill -->
    <surface-form "tokencount="1">kill</surface-form>
  </synset-ref>
  <synset-ref corefcount="2" wid="pwn30:eng-00260470-v" iid="i23019"> <!-- hurt, injure -->
    <surface-form "tokencount="2">injure</surface-form>
  </synset-ref>
  <synset-ref corefcount="2" wid="pwn30:eng-00225150-n" iid=" i36591"> <!-- shooting -->
    <surface-form "tokencount="2">shooting</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-00095280-a" iid=" i500"> <!-- dead -->
    <surface-form "tokencount="1">dead</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-00045888-s" iid=" i233"> <!-- deadly -->
    <surface-form "tokencount="=1">deadly</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-00123783-n" iid=" i36562"> <!-- gunfire, gunshot -->
    <surface-form "tokencount="1">gunfire</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-00220522-n" iid=" i36562"> <!-- murder, slaying, execution -->
    <surface-form "tokencount="1">murder</surface-form>
  </synset-ref>
  <synset-ref corefcount="1" wid="pwn30:eng-00767826-n" iid=" i39445"> <!-- assault -->
    <surface-form "tokencount="1">assault</surface-form>
  </synset-ref>
</ReferenceSet>
```

Figure 1: ReferenceSets for the text fragments referencing the shooter and the event

features, hence exhibit limited variation. Another approach implemented by Vossen and Cybulska (2016), logically matches semantic representations of the *action* mentions, the participants, the time, and the place. Assuming again near-perfect event detection, this approach results in a CoNLL-F1 score of 67.13. For comparison, a baseline system that applies a one-lemma-one-referent heuristics already scores 53.4 CoNLL-F1. As argued in (Cybulska and Vossen, 2014), the ECB+ dataset is very limited in terms of referential variation and within each topic there are only two potential referents to choose between. Concluding, we observe that event coreference systems perform poorly, especially with respect to recall. Applying these corpora to large collections of texts is not likely to give us reliable referential data to derive a ReferenceNet and will not capture sufficient variation. However, the advantage of this approach is that it can be applied to any collection of text.

The **data-to-text** method starts from structured data in which the referents are predefined and searches for texts that make reference to this data, so-called reference texts. Structured event data paired with reference texts appear to exist and are publicly available: GunViolenceArchive (GVA),[3] FireIncidentReports (FR),[4] Railwaysarchive (RA),[5] Gun Violence Database (GVDB),[6] ASN incident database,[7] ASN Wikibase.[8] These resources register event incidents with rich properties such as participants, location, and incident time, and often even provide pointers to one or more reference texts. The number of events and documents is usually high, i.e. there are $\sim$ 9k incidents in RA, and $\sim$ 30k incidents in GVA.

In the *data-to-text* approach, we convert the structured data from such archives to what we call a *microworld*. A microworld is an RDF[9] representation of the referents related to a specific event (e.g. human calamities or economic events) but no more than that. *Reference texts* are then news, blogs, and Wikipedia pages that report on this data. Given the a-priori pairing of microworlds with reference text, we can apply the simple *one-mention-one-referent* principle to obtain reference relations for event mentions for free with a relatively high confidence. By increasingly mixing microworlds and reference texts, we approximate the complexity of reference relations in reality across large

[3] http://gunviolencearchive.org/reports/

[4] https://www.firerescue1.com/incident-reports/

[5] http://www.railwaysarchive.co.uk/eventlisting.php

[6] http://gun-violence.org/

[7] https://aviation-safety.net/database/

[8] https://aviation-safety.net/wikibase/

[9] We use the Simple Event Model (SEM-RDF) to represent events (Van Hage et al., 2011)

volumes of text. By collecting news from different sources on the same or similar events, we approximate true variation in making reference from different perspectives. For example, we can not only take news from different sources with different stances but also vary the time between the event date and the publication date to get articles with different historical perspective. Furthermore, the fact that the data on events from which we start has been created from the perspective of general human interest (e.g. gun violence incident reports) avoids discussion on what establishes an event in text, as we consider only those mentions that directly refer to the reported incident or salient subevents of these incidents.

Although this method may result in more precise reference relations as there is little ambiguity for paired microworlds and reference texts, its downside is the dependency on the availability of the structured data coupled with such reference texts. While for certain types of events such as calamities, sports, and business there may be sufficient data, for others people are less inclined to register events for longer periods. Alternatively, structured event data can also be obtained from DBpedia (Knuth et al., 2015; Elbassuoni et al., 2010), Wikidata (Vrandečić and Krötzsch, 2014), and YAGO2 (Hoffart et al., 2013). As these databases are often linked to Wikipedia articles, references in these articles can be used to find reference texts. Note that we only need the structured data to reconstruct a minimal representation of the referents and we do not need the full representation of the event. Another downside of this approach is that the granularity of the incident is more coarse than the granularity at which the events are reported in the associated news articles. To illustrate this, the GVA collection provides a summary on the incident outcome, whereas the corresponding news documents report on the process that led to this outcome: *firing*, *hitting*, *killing*, *getting injured*, *dying*, etc.

5 The NewsReader event coreference system

We used the NewsReader system (Vossen et al., 2016a) to simulate both a text-to-data and data-to-text process. In both cases, we apply generic semantic parsing to articles, obtaining representations of entities, events, and roles. The output is represented in the Natural Language Annotation Format (NAF) (Fokkens et al., 2014). Coreference for events within a single NAF file is based on a number of steps described in (Vossen and Cybulska, 2016):[10]

1. all predicates from the semantic role layer in NAF are considered as event mentions;
2. we collect all mentions with the same lemma of an SRL predicate throughout the text and consider them to be coreferential;
3. we take the output of WSD for each mention to obtain the best scoring synsets above a threshold. From these synsets, we obtain the highest scoring synsets across all mentions as the most *dominant synsets* for the lemma in the document;
4. we create a coreference set from all the lemma mentions with their dominant senses;
5. all lemma-based coreference sets are compared with each other (cross-lemma) by applying WordNet similarity to the dominant senses across lemma sets
   (a) if their similarity exceeds a preset threshold, we merge the coreference sets across the lemmas aggregating the dominant synsets. In addition, we include the lowest-common-subsumer synset that was responsible for the similarity match.
   (b) if below the threshold, we keep the sets distinct
6. we iterate over the reference sets until there are no changes

For WSD, NewsReader uses the UKB system (Agirre and Soroa, 2009), as well as the supervised It-Makes-Sense system of Zhong and Ng (2010). The output of both systems is used to vote for the most dominant synsets associated with a mention of a predicate. For WordNet similarity, NewsReader uses the WordNet distance measure proposed by Leacock and Chodorow (1998).[11] To be able to capture similarity across nouns and verbs, we extended the WordNet hypernym relations with morphological relations of the type *event* across noun and verb synsets, obtained from the Princeton WordNet website.[12] Below we show two examples of event coreference sets in NAF obtained from two text fragments on the same incident, where the similarity threshold was set to 1.0 and the dominant sense threshold was set to the 80% best-scoring synsets in WSD.

> Curry Bryson , the father of the 11-year - old who police say shot and killed a 3-year - old , appeared in court today for a hearing . ... Barney says it is not the charges against him that have torn his client apart . It is the fact Bryson 's 11-year - old son is accused of shooting and killing 3-year - old Elijah Walker .

```
<coref id="coevent13" type="event">
  <span><target id="t4"/> </span> <!--shooting-->
  <span><target id="t35"/> </span><!--shot-->
  <span><target id="t104"/> </span><!--torn-->
```

[10] Speech acts and so-called grammatical verbs (aspect, auxiliaries) are excluded from this process.

[11] We used the implementation in `https://github.com/cltl/WordnetTools` which allows us to include cross-part-of-speech relations

[12] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

```
5       <exReferences>
6         <exRef conf="1.38" ref="eng-30-02055267-v" source="lcs"/>
7         <exRef conf="0.85" ref="eng-30-01134781-v" source="dom"/>
8         <exRef conf="0.70" ref="eng-30-01597286-v" source="dom"/>
9         <exRef conf="0.74" ref="eng-30-01002740-v" source="dom"/>
10        <exRef conf="0.75" ref="eng-30-02061495-v" source="dom"/>
11        <exRef conf="1.0" ref="eng-30-02484570-v" source="dom"/>
12        <exRef conf="0.72" ref="eng-30-01003249-v" source="dom"/>
13        <exRef conf="0.70" ref="eng-30-02055267-v" source="dom"/>
14        <exRef conf="0.90" ref="eng-30-01137138-v" source="dom"/>
15      </exReferences>
16    </coref>
```

An 11-year - old Detroit boy has been charged with manslaughter in the fatal shooting of 3-year - old Elijah Walker

```
1     <coref id="coevent28" type="event">
2       <span><target id="t148"/></span><!--shooting-->
3       <exReferences>
4         <exRef conf="0.83" ref="eng-30-00225150-n" source="dom"/>
5         <exRef conf="1.0" ref="eng-30-00122661-n" source="dom"/>
6       </exReferences>
7     </coref>
```

In the first fragment, the software lumped together verbal mentions of *shooting*, *shot*, and *torn*. The first two share the same lemma, while they were matched with *torn* through the lowest-common-subsumer (source="lcs") synset `eng3002055267v: buck;charge;shoot_down;shoot;tear`. The similarity score was 1.38. Setting the similarity threshold to 1.5 would prevent merging these mentions. In the second fragment, there is only one mention of the noun *shooting*. We can see that across the documents the verbal and nominal senses will not match on the basis of just the synset identifiers. However, they may still be merged through the form *shooting* or using cross-part-of-speech similarity. From all the mentions, we obtain the most dominant synsets (source="dom") associated by the WSD system according to the threshold setting. The lowest-common-subsumer and the dominant synsets form the basis to compare event coreference sets across documents.

In order to match reference sets across documents, NewsReader first converts NAF representations to SEM-RDF, in which each coreference set represents a unique instance of an event (represented by a unique URI). Each event instance is described with the semantic information associated from all mentions throughout the document. However for the cross-document comparison reported here, we have chosen to match coreference sets only in terms of the overlap of WordNet synsets and surface forms, thus ignoring participants, roles, and temporal relations. The proportion of overlap across instances of events can be set through a parameter. In our experiment, 5% of the synsets or surface forms (in case a lemma has no synsets) need to match for merging instances across different documents.

To simulate the text-to-data approach, *all* the RDF representations of events are compared across *all* the documents. In order to simulate a data-to-text approach, we applied the above cross-document strategy in such a way that events are only compared when the reference texts report on the same incident according to the structured data. This means that *shootings* in documents reporting on different incidents are never compared and cannot constitute coreference relations.

6 Evaluation data

To evaluate both these methodologies, we manually annotated 38 news articles associated with 20 incidents from the GVA data set. The articles were grouped by the incident on which they report together with the structured data on the incidents, e.g. which people got injured or died. We used an annotation schema that differentiates events at different levels of granularity and with respect to the most salient implication derived from the event mention:

incident The incident as a whole is referred to, corresponding to an entry in the structured database.

firing a gun The event of operating a gun without implying somebody got hit.

hit Somebody got hit as a result of shooting without implying death or injury.

miss A gun was used but the bullet missed a person.

injure Somebody got injured as a result of being hit.

die Somebody died as a result of being hit.

For each mention of these events, the annotator creates a unique instance identifier based on the incident, the assigned event type, and the affected victims. When annotating events in documents reporting on the same incident, identity results from same type and victims assigned to mentions whereas non-identity results from a difference in type and/or victim. Documents that report on different incidents never result in identity regardless of the type or victims annotated. Shooting the same person in different incidents is not the same, as well as shooting a different person in the same incident.

The annotation resulted in 138 event instances and 874 mentions in 38 documents. In total, 77 different lemmas were used to make reference to these events. Given these annotations, we can abstract from the instances and group lemmas that make reference to the same *type* of event. Table 1 shows the ReferenceSets derived from the manual annotation for the types of event. Note that the total number of mentions and lemmas is higher because the same word, e.g. *shooting* may occur in multiple reference sets.

Table 1 reveals the large variation based on just 38 documents. We also observe that the event implications follow from very different expressions. For example, *death* can be concluded forward from *fatal shot* or backward from *autopsy*. Especially words making reference to the complete incident show a lot of variation, reflecting different judgments and appraisals.

Table 1: ReferenceSets at the event type level, derived from manual annotation for 20 incidents on gun-violence

Event type	Nr. Variants	Nr. Mentions	ReferenceSets
incident	27	229	accident:39, incident:34, it:34, this:17, murder:15, hunting:14, reckless:14, tragedy:9, happen:8, felony:7, manslaughter:5, what:5, homicide:4, shooting:4, assault:3, case:2, endanger:2, endangerment:2, that:2, violence:2, 's:1, crime:1, event:1, go:1, mistake:1, on:1, situation:1
fire	21	148	shooting:48, fire:25, discharge:16, go:12, shot:9, pull:7, gunman:6, gun:5, gunshot:4, firing:3, shoot:2, turn:2 , accidental:1, act:1, action:1, at:1, handle:1, it:1, return:1, shootout:1, shotgun:1,
hit	11	196	shot:131, discharge:17, shooting:17, strike:16, hit:4, blast:3, victim:3, striking:2, gunshot:1, into:1, turn:1
injure	16	73	wound:36, surgery:13, treat:5, injure:3, stable:3, injurious:2, send:2, bodily:1, critical:1, hit:1, hospitalize:1, hurt:1, injury:1, put:1, stabilize:1, unresponsive:1
die	16	246	death:60, die:52, dead:45, kill:34, fatal:13, lose:9, fatally:7, loss:7, autopsy:6, body:4, take:3, homicide:2, claim:1, deadly:1, life:1, murder:1
Total	114	1043	

7 Evaluation results

We automatically generated ReferenceSets from the 38 annotated documents using the NewsReader pipeline. We used standard settings for dominant-senses (80% top-scoring senses) and similarity (similarity of 2 or higher). Following the methodologies described in section 4, we processed the data twice:

1. **without-i:** comparing all events across all 38 documents, without considering the document-to-incident links from the structured data. This corresponds to the traditional cross-document text-to-data approach.

2. **with-i:** comparing only events across documents if these documents report on the same incident. This method is enriched with data-to-text knowledge.

. In both settings, we only compare event mentions detected by the system and we exclude knowledge about participants, location, and time expressions. We expect *without-i* to lead to more drift in the coreference sets as it will match mentions of events across all documents without the microworld and reference text association. In table 2, we show the coverage results for both, where we make a distinction between the proportion of gold mentions detected and the proportion of gold lemmas. Lemma recall (r) and precision (p) is calculated by comparing the set of lemmas detected by the system to the set of lemmas in the gold annotation. For the mentions evaluation, we compared the frequencies of the lemmas in the texts.

Table 2: Mention and lemma coverage evaluation (r=recall, p=precision, f=harmonic mean) of the NewsReader system output with (with-i) and without (without-i) incident association

	Mentions (874 gold)		Lemmas (77 gold)	
	with-i	without-i	with-i	without-i
r	20.25%	18.19%	49.35%	49.35%
p	59.80%	35.57%	62.30%	46.34%
f	30.26%	24.07%	55.07%	47.80%

We see that *with-i* (incident pairing) performs better in terms of mention recall (+2), precision (+14) and f-score (+6) than *without-i*. For lemma coverage, the recall is the same, but the incident-aware version *with-i* has much higher precision (+16). Overall recall is significantly lower than precision for both methods.

The precision of the data-to-text approach with incident pairing is reasonable (around 60%), though not very high. This can be improved by using better WSD and/or by making the cross-document matching more strict. In the current setting only 5% of the synsets or phrases need to match across documents.

In table 3, we show per event type the ReferenceSets generated by the systems that matched at least one lemma from the gold annotation (the matching lemmas are in bold). We can make a number of observations from these data. First of all, we see that automatic ReferenceSets are more fine-grained than gold sets. This is mainly due to the fact that we use WSD and WordNet similarity to group event mentions in coreference sets. The WordNet synsets and hypernyms do not cover the diverse relations that we annotated for the incidents. Having more relations would merge together reference sets. Furthermore, we see that *with-i* obtains more ReferenceSets, but also more precise ReferenceSets, in comparison to *without-i*.[13] This is to be expected because *with-i* is not allowed to create ReferenceSets across incidents. Finally, we see that multiwords are not considered by NewsReader, which leads to semantic drift for words such as *pull* (the trigger), *take* (a life).

The recall for both methods is really low: around 20% for mentions and 50% for lemmas. Error analysis on the missed recall shows that most of these are not detected as predicates by the semantic role labeler: pronouns (*it*, *this*, *what*), adjectives (*fatal*, *fatally*, *reckless*, *injurious*, *accidental*, *deadly*), and nouns (*dead*, *incident*, *surgery*, *felony*, *autopsy*). Predicate detection is based on the Mate tool (Johansson and Nugues, 2008),

[13]The only exception are the ReferenceSets that include *take*, where the incident pairing generated 4 ReferenceSets and included more wrong mentions than without incident pairing.

Table 3: ReferenceSets at the event type level, derived from automatic annotation for 20 incidents on gun-violence

Type	Reference set with-i	Reference set without-i
textbfaccident:3	**accident:3**	
incident	**act:1** **action:1** **case:3** **crime:1** **happen:14** fact:1 fact:1 **happen:1** **hunting:1** **manslaughter:1** **murder:1** **shootout:1** **tragedy:1** **victim:3**	call:9 make:4 name:2 **act:1 action:1** holler:1 **case:3** **crime:1** **happen:14** occur:2 fact:1 **hunt:2** hunter:1 **hunting:1** **manslaughter:1** **murder:1** **shootout:1** **tragedy:1** **victim:3**
fire gun	**discharge:3** **fire:5** **gun:1** **gunman:1** address:1 deal:1 **handle:1** speak:1 **pull:4** **return:3** **turn:3** **use:2** mother:1	**fire:5 discharge:3** release:3 complete:2 **gun:1** **gunman:1** **pull:4** force:1 **return:3** **turn:3** grow:2 raise:2 mother:6 **use:2** bill:1
hit/fire gun	**shoot:5 shooting:4 shot:2** **shoot:26 shot:7 shooting:4 hit:3 charge:1** **shoot:2 charge:1**	**shoot:23 shot:5 charge:3 hit:3 shooting:2**
hit	**hit:3 shoot:3** **strike:2**	 **strike:2**
injury	**send:7** post:4 message:1 message:1 **send:1** **treat:2** **wound:3** **hurt:3**	**send:6** carry:5 post:5 letter:1 message:1 transport:1 handling:3 **treat:2** deal:1 **handle:1** manage:1 **wound:3** **hurt:3** back:2 suffer:2 support:2
die	**death:7 die:4** **die:9 death:7** run:1 **kill:12** house:2 live:2 **life:1** **life:8** live:5 house:2 **lose:4** **loss:1** **put:4** place:1 place:1 **put:1** say:52 **take:21** involve:10 need:9 come:8 get:8 tell:3 ask:2 bring:2 carry:2 want:2 conduct:1 involve:10 come:8 get:8 **take:4** need:3 bring:2 want:2 carry:2 **take:2** ask:2 **take:2** need:1	**die:9 death:5** run:5 **kill:12** family:13 **life:7** home:5 live:4 house:1 **lose:4 loss:1** **put:4** place:2 set:2 lay:1 say:146 tell:17 **take:14** involve:7 need:6 ask:5 conduct:3 statement:2 want:2 bring:1

which is trained on PropBank (Kingsbury and Palmer, 2002), and NomBank (Meyers et al., 2004). Improving the recall for lexical coverage therefore primarily requires improving the coverage of these resources.

8 Conclusions and future work

In this paper, we present ReferenceNet: a network of referential relations between synsets that is complementary to WordNet and word embeddings. ReferenceNet consists of ReferenceSets that group synsets and words that make reference to similar entities and events within similar topical contexts. Typically, ReferenceSets reflect different local contexts and perspectives within a shared topical context as opposed to synsets and word embeddings that capture similar local contexts. We described two methods to derive ReferenceSets from textual data. We evaluated the approaches against a manually annotated data set. We concluded that precision is reasonable, whereas recall is low, mainly due to poor recall of predicates. We also observed that coreference relations are missed because WordNet does not sufficiently capture coherence and perspective relations, resulting in smaller ReferenceSets. The evaluation further showed that ReferenceSets created with a data-to-text approach have higher recall and precision. In future work, we want to capture more referential variation. Event coreference can be improved using other coherence measures, especially when comparing coreference sets across documents. The fact that WSD already restricts the association of concepts by part-of-speech limits the matching in the current system. We will also extend to other types of events and contexts. Finally, entity coreference can be included by exploiting semantic matches of noun phrases and semantic roles.

Acknowledgements

The work presented in this paper was funded by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines". We also thank the reviewers for their constructive comments.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

C. F. Baker, C. J. Fillmore, and B. Cronin. 2003. The structure of the framenet database.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422. Association for Computational Linguistics.

Reinhard Blutner. 1998. Lexical pragmatics. *Journal of semantics*, 15(2):115–162.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*.

Z. Chen and H. Ji. 2009. Graph-based event coreference resolution. pages 54–57.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552.

Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, and Gerhard Weikum. 2010. Searching rdf graphs with sparql and keywords. *IEEE Data Eng. Bull.*, 33(1):16–24.

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

G. Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophischen Kritik*, 100:25–50.

H. P. Grice, P. Cole, and J. Morgan. 1975. Logic and conversation. pages 41–58.

Nicola Guarino. 1999. The role of identity conditions in ontology design.

J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28???61.

E. Hovy, T. Mitamura, F. Verdejo, J. Araki, and A. Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187. Association for Computational Linguistics.

P. Kingsbury and M. Palmer. 2002. From treebank to propbank. pages 1989–1993.

Magnus Knuth, Jens Lehmann, Dimitris Kontokostas, Thomas Steiner, and Harald Sack. 2015. The dbpedia events dataset. In *International Semantic Web Conference (Posters & Demos)*.

Saul A Kripke. 1972. Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

S. Levinson. 1983. *Pragmatics*. Cambridge University Press.

A. Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June.

Yo Matsumoto. 1995. The conversational condition on horn scales. *Linguistics and philosophy*, 18(1):21–60.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.

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.

Hilary Putnam. 1973. Meaning and reference. *The journal of philosophy*, 70(19):699–711.

W Quine and O Van. 1960. Word and object: An inquiry into the linguistic mechanisms of objective reference.

Erich H Rast et al. 2007. *Reference and indexicality*. Logos-Verlag.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136.

Piek Vossen and Agata Cybulska. 2016. Identity and granularity of events in text.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016a. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

Piek Vossen, Francis Bond, and J McCrae. 2016b. Toward a truly multilingual global wordnet grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Edda Weigand. 1998. *Contrastive lexical semantics*, volume 171. John Benjamins Publishing.

Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *arXiv preprint arXiv:1504.05929*.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.