A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts

Martin Scholz

Universität Erlangen-Nürnberg,
Erlangen, Germany
martin.scholz@fau.de

Abstract. The detection of event mentions in free text is a key to a deeper automatic understanding of the text's contents. In this paper we present ongoing work on mechanisms to detect events in German texts in the domain of cultural heritage documentation. A central role plays a hand-crafted mapping of CIDOC CRM[1] events to GermaNet synsets to ease the process of creating a lexicon for automatic event detection. We discuss two approaches and insights gained from the mapping process and correct modelling of event mentions.

1 Introduction

In cultural heritage, free text is an important source of information and a popular form of documentation. For the latter, free text is often combined with structured metadata records. While the records provide basic, standardized metadata, the texts contain more detailed descriptions or additional information. Structured metadata can be accessed and processed quite well by machines For the contents of free text, however, this does not hold. Although there exist various methods for automatic information extraction, currently none can reach the high quality of expert-proven data necessary for academic research. Their efficacy varies heavily with text properties such as language, genre, etc; and this most likely will not change in near future. It is therefore desirable to semantically enrich texts with human revised annotations in order to extract its contents in a machine-processable way with quality sufficient for scolarly research.
One approach is to assist human annotators with automatic text analysis methods, providing for annotation proposals. Such an approach is implemented in the WissKI system, as described in Sect. 3.3.
Basically, such detection algorithms rely on one of two types of data resources for computing their heuristics: Either on a large-scale annotated corpus or on a (hand-made) lexicon. For common named entity classes like persons, places, organisations and times there are hand-annotated corpora and ready-to-use automatic annotation tools available, although languages other than English are supported much more rarely [14], [4], [3]. Events[2] are covered less frequently. The

[1] In this paper we always refer to version 5.0.4 of the CIDOC CRM [2].
[2] Note that the notion of what the term "event" means varies in information retrieval. E.g. some literature focuses rather on (historical) periods like "industrialisation". In this paper we align our understanding of the term with the class *E5 Event* in the CRM.

Timebank corpus[3] [8], an English corpus annotated with TimeML[4] [1] mark-up language, also contains annotations of events and there is some literature about event detection [9]; again, mostly for English. For our target language, German, we are currently not aware of any freely available corpus with event annotations or tools for automatic event detection.

In this paper we describe a mapping of CIDOC CRM event classes to GermaNet, a wordnet for the German language. From the mapping and GermaNet, a word list can be compiled that is the basis for an event detection algorithm. As we are not aware of available German corpora tagged with CIDOC CRM event classes, we also built a small manually annotated corpus of text from museum documentation, which we use for development and evaluation.

The rest of the paper is structured as follows: First, the lexical resource for our mapping, GermaNet, is briefly described. In the following section we present a simple and a more elaborate mapping strategy and shortly discuss their strengths and weaknesses. Then, the detection algorithm is described and evaluated against a small hand-crafted corpus. Further, we describe its application in the WissKI system. In Sect. 4 observations and challenges for future work are discussed. Finally, we conclude with Sect. 5.

1.1 German Wordnet

GermaNet[5] [6] is a German wordnet. Its structure is based on the Princeton WordNet[6] for English. Unlike Princeton WordNet, GermaNet is not open data, but only free for academic research. The work described here is based on GermaNet version 7.0.

Key concept of the family of wordnets is the so-called synset, a set of words[7] that are synonyms in a certain textual context. A synset is thus an equivalence class, i.e. the words of a synset can be used interchangeably in that context.

A word can participate in several synsets, reflecting large or small shifts in its meaning. The meanings of a word are numbered, so that a specific meaning — a so-called word sense — can be identified by the word and an integer. Likewise, a synset can be identified by the word sense of one of the words it contains. GermaNet distinguishes three parts of speech or word categories: noun, verb, and adjective.

Synsets are linked to each other by certain semantic relationships, like antonymy or meronymy. The predominant one is hypernymy. Synsets are usually arranged hierarchically according to the hypernym-hyponym relation, forming a thesaurus. A synset may have multiple hypernyms.

A synset can be regarded as resembling the common meaning of a set of words. Thus, a synset can be seen as the lexical equivalent to a concept in an ontology while the hypernymic relation corresponds to the subclass relation. In fact, there have been some proposals to model wordnets as ontologies [7].

---

[3] http://www.timeml.org/site/timebank/timebank.html

[4] TimeML is a vocabulary for annotating temporal expressions in text. See http://www.timeml.org

[5] http://www.sfs.uni-tuebingen.de/lsd

[6] http://wordnet.princeton.edu

[7] Strictly speaking, a synset contains one or more so-called lexical units. A lexical unit contains the uninflected word form with possible orthographic variants. To keep it simple, we will not distinguish "word" from lexical unit.

2 The Mapping Mechanism

The idea of using GermaNet for event detection is that the structure of GermaNet can be exploited to generate large lists of words identifying CRM events by mapping an event class to a handful of synsets, rather than generating a list of words by hand. We assume that if the words of a synset can be used to denote a CRM event class, then its hyponyms are likely to also support this class. The more hyponyms the synset has, the more words can be selected with relatively small effort.

In this section we present two approaches for such a mapping[8] for CRM *E5 Event* and its subclasses, with two exceptions: *E13 Attribute Assignment* and its subclasses were not taken into account, as a first examination of the corpus data indicated that instances of this class are preferably expressed grammatically differently from other event classes. This may be due to the generic, metalevel-like nature of *E13 Attribute Assignment*. *E87 Curation Activity* was excluded primarily as it was out of scope of our research, but also because we were unsure about its extent and what words support it.

2.1 A Simple Approach

We first implemented a naive mapping approach. For each event class a small set of synsets was determined with two conditions:

1. the synset supports the concept
2. all hypernymic synsets do not support the concept

A synset supports a concept if one of its word senses refers to the class. Note that it is not required that each word sense of a word must refer to the class. Figurative use of words was not taken into account.

The second condition brings about that only the topmost synsets (in the sense of hypernymy) relatable to that event class are chosen, leading to a minimal set of synsets.

With appropriate tools for exploring the synset graph like GermaNet Explorer[9] such a mapping was built quite rapidly.

The mapping rules are expressed in XML:

```
<class name="ecrm:E67_Birth">
  <synset pos="v" word="gebären" sense="1" />
  <synset pos="n" word="Geburt" sense="1" />
  <synset pos="n" word="Geburt" sense="2" />
  <synset pos="n" word="Geburt" sense="3" />
</class>
```

Fig. 1. Declaration of synsets mapping to the *E67 Birth* event

A conversion programme was developed that compiles the synsets to a list of words: First, all hyponymic synsets are fetched from GermaNet. Then, the words contained in the synsets are extracted and printed with their word category. Duplicates are omitted. The result is again an XML mapping of event classes to words as shown in Fig. 2.

---

[8] The second mapping approach is available as an XML file for downloaded from http://wiss-ki.eu/node/167.

[9] http://www.sfs.uni-tuebingen.de/lsd/tools.shtml#GermaNet-Explorer

```
<class name="ecrm:E67_Birth">
    <word lemma="gebären" pos="v"/>
    <word lemma="entbinden" pos="v"/>
    <word lemma="niederkommen" pos="v"/>
    <word lemma="werfen" pos="v"/>
    <word lemma="laichen" pos="v"/>
    ...
    <word lemma="Geburt" pos="n"/>
    <word lemma="Drillingsgeburt" pos="n"/>
    <word lemma="Niederkunft" pos="n"/>
    <word lemma="Entbindung" pos="n"/>
    <word lemma="Totgeburt" pos="n"/>
    ...
</class>
```

Fig. 2. Excerpt from the compiled word list for *E67 Birth*

2.2 Problems of the First Approach

This simple approach shows two shortcomings:
The first problem arises from the polysemy of words. A word with different meanings — and thus contained in different synsets — is less likely to actually denote a specific event class than a word with only one meaning. Also, one meaning might be more frequent than another.
The predominant problem with this first approach, however, is that the scope of a CIDOC CRM event and the meaning of GermaNet word senses and synsets virtually never match exactly, but rather overlap. So, although a synset may support a CRM event class, the words of an hyponymic synset, however, may in no case support the event. This is illustrated by two prominent cases:
In CRM, the *E67 Birth* event only holds for humans. The birth of other living beings like animals is modelled with *E63 Beginning of Existence*. The top synset "gebären" in Fig. 1 supports the notion of a human birth and its words are the most commonly used in German for such an event. But they also may denote an animal birth. Consequently, some lower synsets introduce words that cannot be applied (unless as a colloquial or pejorative term) to human births, like "werfen" (mostly used for mammals with a bunch of offspring) or "laichen" ("spawn")[10] as shown in Fig. 2.
Another special case arises from the CRM clearly dividing things into material (*E19 Physical Thing*) and immaterial (*E28 Conceptual Object* and *E90 Symbolic Object*). This also affects the CRM event classes, as there are different classes for both branches: e.g. *E12 Production*/*E11 Modification* vs. *E65 Creation*. The German language and thus GermaNet, however, do not reflect this division. As a result, it is hardly impossible to find sufficiently broad synsets for which all words and hyponyms support the event. Only synsets with specialized meaning and with no or very little hyponyms fulfill this criterion. Synsets with frequently used words like "erschaffen", "erzeugen", "produzieren" (create, produce) all contain a wild mixture of hyponymic synsets applicable to events affecting either material things or immaterial things or both.

[10] In some cases GermaNet seems to be inconsistent: While "werfen" and "laichen" are grouped as birth, bird reproduction words like "legen" (lay an egg) or "schlüpfen" (hatch) are not.

An option would be to change the policy described in the previous section and only select synsets with words which always imply the event class. However, this leads to significantly less synsets and often excludes the most commonly used words, like “gebären” from *E67 Birth*.

2.3 A more fine-grained mapping

To overcome the shortcomings of the first approach, the mapping was extended so that hyponymic synsets can be excluded from the compilation process. For the XML notation, two modes were defined:

1. The element `<exclude_synset>` references a single synset that will be excluded. Its descendants are also excluded unless they can be reached via another branch or by another selected synset.
2. The boolean attribute `descend` for the `<synset>` element controls whether hyponyms should generally be included or excluded for this very synset. If set to `false`, all hyponyms of a synset are excluded by default.

The latter is primarily for convenience. However, it can also be regarded to lower the degree of semantic overlap of the synset and the CRM class: If set to `true`, the overlap is deemed to be rather high, as hyponyms are included by default. Analoguously, when `false`, the overlap is rather low.

Sometimes, synsets should be included that were implicitly excluded by one of the two methods. In such a case, the synset is explicitly selected, i.e. added to the synset list just like a top synset. Fig. 3 shows two examples: The Birth event now excludes all verbs denoting animal reproduction. The *E66 Formation* event is mapped to the synset “Heirat” (wedding) which mainly contains other activities as hyponyms like wedding anniversaries that don’t support *E66 Formation*. Therefore, they are excluded by default. The hyponym “Liebesheirat” (“marriage for love”), however, is explicitly included.

```
<class name="ecrm:E67_Birth">
  <synset pos="v" word="gebären" sense="1">
    <exclude_synset word="werfen" sense="5" />
  </synset>
  ...
</class>

<class name="ecrm:E66_Formation">
  ...
  <synset pos="n" word="Heirat" sense="1" descend="false" />
  <synset pos="n" word="Liebesheirat" sense="1" />
  ...
</class>
```

Fig. 3. Synsets can be explicitly excluded from the mapping

Although the events affecting material or immaterial things can be mapped quite accurately, the mapping is still not optimal as a lot of excludes have to be defined: The *E11 Modification* event maps to five topmost synsets, but with about 200 exclude statements. In such cases, the mapping process becomes quite time-costly and error-prone as the whole subtree must be scanned for synsets to exclude.

The conversion tool was adapted accordingly. Furthermore, each word will be given a confidence value between 0 and 1 that resembles the confidence that the intended meaning or word sense of the word in the given context is one of the word senses denoting the event. It is computed as follows:

$$confidence(w) = \frac{s_{w,e}}{s_w}$$

$s_{w,e}$ is the number of word senses of word w contained in the mapping for event e and s_w is the total number of word senses for word w.
The confidence can be used by a parser to rank event findings. However, this value only very roughly approximates the actual frequency of word senses in human language or a corpus.[11]

3 Event Detection

The compiled word lists are used for list-based event detection in the cultural heritage domain. The texts are tokenized, lemmatized and tagged with parts of speech (POS) using the Stuttgart TreeTagger [11]. A small script resolves separable verb particles, i.e. adds a particle to the corresponding verb lemma. [12]
In order for a token to be annotated as denoting an event, its lemma must occur in the corresponding word list and the POS tag must match the word category. Tokens may be annotated with multiple event classes. However, only the most specialized classes are kept, i.e. if a token is annotated with *E9 Move* and *E7 Activity*, the latter one is discarded as it is implicit in the former one.
At the moment, the algorithm does not perform word sense disambiguation. Words are annotated with possible events for each word sense. However, event annotations can be ranked according to the confidence value mentioned above.

3.1 Light Verbs

In German, light verb constructions are frequent, especially in scientific writing. Light verb constructions consist of a verb and a noun phrase, usually a nominalized verb, sometimes also including a preposition. Within this construct the noun carries the overall meaning, while the verb is reduced to only add a certain aspect[13] like causation. Typical examples include “erfolgen” or “stattfinden” (“take place”) together with an event-baring noun and rather fixed or lexicalized collocations like “zum Einsturz bringen” (“cause to collapse”).
A lot of light verbs can also occur on their own with a distinguished meaning (e.g. “bringen" then meaning “to bring”) and as such may also denote an event.

[11] This could be done in a further step, though, by computing the word sense frequencies from a corpus annotated with word senses, like the WebCAGe corpus (http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/webcage.html).

[12] In German language, a separable verb particle is a part of a verb that may occur separated from the verb stem in a proposition. The particles usually change the meaning of the verb, leading to totally different event classes.

[13] In German linguistics the common term is Aktionsart. A light verb usually shifts the focus to a certain aspect of the event, like beginning, end, result or cause.

In contrast, a light verb does not denote an event. A parser ignorant to light verb constructions will therefore produce much more false positives.
We also included a lexicon-based postprocessor to detect light verb constructions. Our parser uses a small hand-crafted lexicon and a dependency parser[14] in order to find such constructions. For a match, the verb is stripped off any event annotations. Event annotations for the noun part are augmented with aspect information provided by the light verb. We expect the aspect information to be a valuable hint in the role labeling phase that we plan to implement and for the right event modeling (see Sect. 4.2).

3.2 Evaluation on a Small Annotated Corpus

The coverage of the mapping was tested on a small corpus of short texts about museum objects.[15] The texts were annotated with event mentions manually.
Currently, the corpus contains 50 annotated texts with over 3000 tokens and 500 annotations.
For evaluation, a found annotation would be considered relevant if the corpus contained an annotation with same event class and that had at least 50% overlap. Conversely, a relevant annotation would be marked as missed, if the parser's output would not contain an annotation that suffices these conditions.
We achieve a precision of 59% and recall of 72%.

3.3 Use in the WissKI System

Our event detection system is developed as a part of the WissKI[16] virtual research environment[17]. WissKI is web-based, extending the popular content management system Drupal. It consistently relies on semantic web technology. Data is stored according to the CIDOC CRM in its OWL-DL implementation Erlangen CRM[18].
In WissKI, one form of data acquisition consists of semantically annotating free text in a WYSIWYG editor [12], [5]. From the enriched text, RDF triples can then be generated automatically. Annotations include entities like persons, objects, places, calendar dates, and events, and relations between these entities. The annotation process is designed to be semi-automatically:[19] WissKI provides the user with multiple annotation proposals. The user may always edit machine-produced annotations. Thus it is more important for the system to compute a (ranked) list of possible annotations than a single best solution. From this follows immediately that a higher recall is more favourable than high precision.

---

[14] We use the dependency parser ParZu [13] from the University of Zürich http://kitt.cl.uzh.ch/kitt/parzu/.

[15] The texts describe European works of art and are part of the online presentation of the exhibition about Renaissance, Baroque, and the Age of Enlightenment by the Germanic National Museum, Nuremberg, Germany.

[16] The WissKI project was funded by the German Research Foundation (DFG) from 2009-2012. Since then the WissKI software has been further developed.

[17] http://wiss-ki.eu

[18] http://erlangen-crm.org

[19] We don't expect natural language processing techniques to become accurate enough to obtain high-quality annotations in the near future. Therefore, machine-generated annotations must be approved by human experts to garantee annotation quality that meets academic standards.

4 Further challenges

The work on CRM event mapping and detection has raised some issues that we want to address in the future.

4.1 Mapping to English Wordnet

For English, there are much more sources of annotated data, but also linguistic resources and tools for event detection than for German. Consequently, a similar mapping for the English Princeton WordNet could reveal interesting insights for event detection, also for German. The Interlingual Index[20], an outcome of the EuroWordNet project, serves to build mappings between wordnets of various languages by introducing an intermediate layer. The mapping between GermaNet and Princeton Wordnet is kept up-to-date by the makers of GermaNet.[21] It remains to be seen if it could serve as a starting point or it is better to start from scratch.

4.2 When is an event a CRM event and of what kind?

The detection of events is just a first step towards an accurate modelling of events according to the CIDOC CRM. In fact, an event annotation can be modelled quite differently in CRM, depending on the context:
The CRM only models events as *E5 Event* if they actually took place. Hypothetical events, instead, should be modelled as conceptual objects like *E55 Type* or *E29 Design or Procedure*.
Further, a word literally denoting a certain event class may be actually modelled as a superclass of the event. For example, this is the case for events expressed with words that usually denote specializations of *E7 Activity* like *E12 Production* or *E8 Acquisition*, but that have been interrupted and produced no result. An example from the corpus is
"[...] Dentatus weist die Geschenke [...] zurück."
"[...] Dentatus rejects the presents [...]"
where the implied transfer of ownership (to give a present) could not be completed, and thus is just an *E7 Activity*. Nonetheless, it is of importance to also model the intended action. Likewise, events normally supporting (sub)classes of *E63 Beginning of Existence* or *E64 End of Existence* may fall back to *E5 Event*.
It is also important to detect how many event instances a word evokes. Based on the data in the corpus, we differentiate three cases depending on the number of individual events that are referred to:

individual: the word refers to only one single event instance. In most cases this event can be modelled as CRM event, unless it is hypothetical.

collection: the word refers to multiple but distinguished event instances of the same class. As in the case of an individual the events can be modelled as CRM events.

class: the word refers to a class of events rather than to event instances. Often, processes are described and so appropriate CRM classes would be *E29 Design or Procedure* or similar — as with hypothetical events.

[20] http://www.illc.uva.nl/EuroWordNet/
[21] http://www.sfs.uni-tuebingen.de/lsd/ili.shtml

The border between collection and class can be blurred and hard to identify. A collection of events is usually linked to a description of a well-defined collection of items or group of people. A class usually co-occurs with terms denoting classes of items. Thus, the correct modelling of events is highly dependent of the entities in context.
For the right modelling grammatical numerus is an important clue. The singular invokes the individual case while the plural invokes the collection or class case. Also, key words like "solche" ("such"), "diese" ("these") and other determiners can help to distinguish a class from a collection.[22]
TimeML also addresses this issue by distinguishing between event tokens and event instances, for a collection or individual. Classes (called "generics"), however, are not treated by TimeML [1], [10, pp. 1–8, 32–35].

4.3 Implicit Events

As seen in the sentence "Dentatus rejects the presents" an event mention can be co-triggered by a word primarily referring to an object or person; in this case the word "presents", denoting the material things in first place, but also the mode of handing over. Other frequent words include "Maler" (painter), "Gemälde" (painting) and family relations like "Tochter" (daughter) or "Vater" (father), including a *E12 Production*, *E65 Creation* or *E67 Birth* event, respectively.
It is hard to draw a line if event classes should be co-triggered with a certain word and if so, which ones. While the aforementioned "Gemälde" clearly triggers an *E12 Production*, it is not so clear for "Kunstwerk" (*work* of art) and "Objekt" (object) would not — although "Gemälde" and "Objekt" are both hyponyms of "Kunstwerk".
We have no clear guidelines yet. Our current practice is that a word denotes an event if it was somehow morphologically derived from a word denoting that event.
Nevertheless, such information can help in finding the right relation for constructions like
"Albrecht Dürer's painting"
"Albrecht Dürer's house"
In the first phrase, the production event implied in "painting" favours this event as link between the two entities. In contrast, in the second phrase, the default possession or ownership relation is more likely to be meant.

5 Conclusion

We presented a partial mapping of CRM event classes to GermaNet, a German wordnet. The mapping is used as a lexicon for detecting event mentions in free text. The mapping does not claim to be complete and will be refined in the future while applied to more textual sources and other cultural heritage domains. Likewise, we will extend the algorithms and tools for event detection so that they better suit the needs of the users.

[22] In fact, determiners have a long history in linguistics of functioning as a discriminator for class or instance.

Acknowledgement

We are very thankful to Guenther Goerz for helpful suggestions and discussions and to the reviewers for valuable hints and suggestions.

References

1. TimeML 1.2.1. a formal specification language for events and temporal expressions (October 2005)
2. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.e.: Definition of the CIDOC Conceptual Reference Model — Version 5.0.4
3. Faruqui, M., Padó, S.: Training and evaluating a german named entity recognizer with semantic generalization. In: Proceedings of KONVENS 2010. Saarbrücken, Germany (2010)
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). pp. 363–370 (2005)
5. Goerz, G., Scholz, M.: Adaptation of NLP techniques to cultural heritage research and documentation. Journal of Computing and Information Technology 18 (2010), http://cit.srce.hr/index.php/CIT/article/view/1918
6. Kunze, C., Lemnitzer, L.: GermaNet - representation, visualization, application. In: Proceedings of LREC 2002. pp. 1485–1491 (2002)
7. Kunze, C., Lemnitzer, L., Lüngen, H., Storrer, A.: Modellierung und Integration von Wortnetzen und Domänenontologien in OWL am Beispiel von GermaNet und TermNet. In: Proceedings of KONVENS 2006. pp. 91–96. University of Konstanz (2006)
8. Pustejovsky, J., Hanks, P., Saurí, R., See, Andrew Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M.: The TIMEBANK Corpus. In: Proceedings of Corpus Linguistics. pp. 647–656 (2003)
9. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: A Robust Event Recognizer for QA Systems. In: Proceedings of HLT/EMNLP. pp. 700–707 (2005)
10. Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines Version 1.2.1 (January 2006), http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf
11. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland (1995)
12. Scholz, M., Goerz, G.: Wisski: A virtual research environment for cultural heritage. In: Raedt, L.D., Bessière, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P.J.F. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 1017–1018. IOS Press (2012)
13. Sennrich, R., Volk, M., Schneider, G.: Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. Hissar, Bulgaria (2013)
14. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of CoNLL-2003. pp. 142–147. Edmonton, Canada (2003)