Historical Event Extraction from Text

Agata Cybulska
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
ak.cybulska@let.vu.nl

Piek Vossen
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
p.vossen@let.vu.nl

Abstract

In this paper, we report on how historical events are extracted from text within the Semantics of History research project. The project aims at the creation of resources for a historical information retrieval system that can handle the time-based dynamics and varying perspectives of Dutch historical archives. The historical event extraction module will be used for museum collections, allowing users to search for exhibits related to particular historical events or actors within time periods and geographic areas, extracted from accompanying text. We present here the methodology and tools used for the purpose of historical event extraction alongside with the first evaluation results.

1 Introduction

The research project Semantics of History[1] is concerned with the development of a historical ontology and a lexicon that will be used in a new type of information retrieval system. In historical texts the reality changes over time (Ide & Woolner, 2007). Furthermore, historical realities can be seen differently depending on the subjective view of the writer. In the design of our search system, we will take into consideration the change of reality and the diverse attitudes of writers towards historical events so that they both can be used for the purpose of historical information retrieval.

In the first phase of the project we researched how descriptions of historical events are realized in different types of text and what the implications are for historical information retrieval. Different historical perspectives of writers correspond with genre distinctions and correlate with variation in language use. Texts, written shortly after an event happened, use more specific and uniquely occurring event descriptions than texts describing the same events but written from a longer time perspective. Statistical analysis performed within the first phase of the project confirmed this hypothesis[2]. To capture differences between event representations and to identify relations between historical events, we defined a historical event model which consists of 4 slots: a location slot, time, participant and an action slot (see also Van Hage et al 2011 for the formal SEM model).

After arriving at an understanding of how to model historical events, we moved on to actually extracting events from text. In this paper we report on our approach into historical event extraction from textual data about the Srebrenica Massacre from July 1995[3]. There are two problems that had to be tackled for the purpose of this task: 1) extraction of event actions with their participants, locations and time markers and 2) filtering of events lacking historical value from all events extracted by the system. We believe that event actions and their participants, locations and time markers can be extracted based on some syntactic clues, PoS, lemma and combinatory information together with semantic class definition and exclusion by means of Wordnet. Historical filtering can be performed through semantic classification of event actions.

[1] The Semantics of History is funded by the Interfaculty Research Institute CAMeRA at the Free University Amsterdam as a collaboration of the Faculties of Arts and Exact Science: http://www2.let.vu.nl/oz/cltl/semhis/index.html.

[2] For details see Cybulska, Vossen, LREC 2010.

[3] The Srebrenica corpus consists of 78 Dutch texts. For more information on the design of the corpus see Cybulska, Vossen (2010).

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 39–43,
Portland, OR, USA, 24 June 2011.

We tested this hypothesis within the KYOTO framework[4].

2 Related Work

Two other projects concerned with extraction of historical information are the FDR/Pearl Harbor project and the New Web Portal. The latter[5] aimed at creation of a digital archive of historical newspapers of the National Library of Finland[6]. Within the project a semantic search system for historical texts was created using a common ontology with semantically annotated cultural objects (Ahonen and Hyv¨onen, 2009). Related content is being linked through semantic annotation of historical texts based on ontology labels which presupposes that only high level historical events from text were annotated. The Pearl Harbor project aimed at facilitating enhanced search and retrieval from a set of documents from the FDRL library by utilizing a series of multiple temporally contextualized snapshot ontologies determined by the occurrence of key historical events (Ide & Woolner, 2007). We did not manage to find evaluation results for any of the two projects. Traditional approaches to event extraction that do report evaluation results use models that severely restrict the relations. They achieve high precision but poorly represent the text as a whole. E.g., Xu et. al. (2006) report over 80% precision for prize award extraction and Tanev et. al. (2008) 74% precision for violent events and disasters. Our approach models more events in a text and events of a broader scope, more comparable to Wunderwald (2011), who extracts participants and roles from news in general, reporting 50-60% precision. Wunderwald uses a machine-learning approach, while our method is knowledge-based. Furthermore, Wunderwald does not distinguish historical from non-historical events.

3 Historical Event Extraction

3.1 Generic Event Extraction by means of KYOTO

KYOTO tools were specifically designed to extract events from text. This pipeline-architecture of linguistic processors generates a uniform semantic representation of text in the so-called Kyoto Annotation Format (KAF)[7]. KAF is a stand-off format that distinguishes separate layers for text tokens, text terms, constituents and dependencies. It can be used to represent event actions with their participants, locations and time markers. For the purpose of this research, the Srebrenica corpus was processed by means of the KYOTO – architecture. First, the corpus was tagged with PoS- information; it was lemmatized and syntactically parsed by means of a dependency parser for Dutch - Alpino[8]. Next, word sense disambiguation was performed[9] and the corpus was semantically annotated with labels from the Dutch Wordnet[10] and ontological classes. Generic event information stored in the KAF – format can be extracted within KYOTO by means of Kybot-profiles which are stored in the XML format[11]. These profiles define patterns over different layers in KAF and create a semantic output layer for matches over these layers.

3.2 Semantic Tagging of Historical Events

To extract historical events we developed 'historical' Kybot-profiles which define appropriate constructions and semantic classes of historical actions and their participants, locations and time markers. In these profiles, the semantic action classes are used to distinguish historical from non-historical events. The semantic type specification was derived from manual tagging of historical event slots by means of the KAF-annotator[12] in 5 development texts from the Srebrenica corpus[13]. Manually tagged historical event actions as well as participants, locations and time markers were automatically mapped with corresponding Wordnet synsets. In case of multiple senses assigned per word the appropriate Wordnet ID was manually chosen.

Historical event tagging with Wordnet ID's revealed a few problematic issues. For a number of

[4] For more information about the KYOTO - project (www.kyoto-project.eu) see Vossen et al (2008a).
[5] The New Web Portal is part of the National Semantic Web 2.0 (FinnONTO 2.0) project.
[6] http://digi.lib.helsinki.fi/sanomalehti/secure/main.html
[7] Kyoto Annotation Format is described in Bosma et al (2009).
[8] http://www.let.rug.nl/vannoord/alp/Alpino/
[9] For word sense disambiguation the UKB system (http://ixa2.si.ehu.es/ukb/) was used. For more information the reader is referred to Agirre & Soroa (2009).
[10] For more information see Vossen et al (2008b).
[11] For more information see KYOTO deliverable 5.4 at http://www.kyoto-project.eu/.
[12] See tools at http://www.kyoto-project.eu/.
[13] The development set contains one Wikipedia entry, two educational texts and two newspaper articles written a few years after the Srebrenica massacre happened.

locations, time markers, participants and actions there were no Wordnet synsets automatically assigned. No WN-concepts were found for geographical names as *Srebrenica* or *Zagreb*. Also person and organization names (*Mladic, Dutchbat III, NIOD*) and dates would not get any synsets assigned. The same applies to compounds (*moslimmannen* 'Muslim men', *VN-militairen* 'UN soldiers'), pronoun participants and loanwords: (such as *safe haven* in a Dutch text). Furthermore there were some historical senses missing in the Dutch Wordnet (such as *vredesoperatie* 'peacekeeping operation', *oorlogspad* 'warpath'). To be able to handle proper names we used a named entity recognition module. By means of NER we added dates and geographical names to KAF so that we could further use them for the extraction of time markers and locations. In the future, we will look into compound splitting and we are also going to add the missing historical senses to the Wordnet database.

After identifying historical WN-synsets, we automatically determined the most informative hypernyms of the seed terms per historical label. Based on the chosen hypernyms (and their hyponyms), we manually selected a number of semantic classes to be able to identify event locations, time markers, participants and historical actions in historical texts. We defined six semantic classes denoting: human participants, time periods, moments in time, places, historical and motion actions. Furthermore we specified six more action classes to filter out non historical and potential events: actions indicating modality, polarity, intention, subjectivity, cognitive (also rarely of historical importance) and contentless actions. Next, we derived a table that assigns one of the ontological classes to every synset in Wordnet on the basis of the relations to the labeled hypernyms. All KAF-files were then annotated with the twelve semantic classes, on the basis of the Wordnet synsets assigned by the WSD module and this mapping table.

4 Kybot Profiles

Kyoto-Kybot extracts events from KAF by means of Kybot profiles. Based on event descriptions from the development set 402 profiles were defined, using semantic and constructional information and specifically PoS, lemma, compositional and semantic restrictions with regards to locations, time expressions, event actions and participants.

The current version of the system uses 22 profiles to extract historical actions, based on semantic tagging by means of Wordnet and the specification of some compositional properties. Historical actions are the most significant part of historical event extraction. They serve to distinguish historical actions from the non-historical ones and to identify parts of the same historical event. The profiles extract both, verbal actions (such as *deport, murder, occupy*) and nominal ones (such as *fight, war* and *offensive*) as well as actions with a syntactic object (*sign a treaty, start the offensive* etc). Next to the semantic class of historical actions also motion actions (often occurring with a goal or result phrase as *transport into* a location) are extracted as potential historical event actions. The action profiles exclude from the output the non-historical semantic action classes and by that the non historical events are filtered out.

For the extraction of historical participants we now use 314 profiles. The variation within historical participant descriptions of the development set was, as expected, much higher than the diversity of formulations denoting other event parts. Participant profiles specify noun phrases (also proper names) organized around the semantic class of human participants[14]. It is a relatively common phenomenon in historical event descriptions that geographical proper names are used for referral to participants. So we also created some profiles identifying country and city names occurring in the subject position of active sentences.

To extract historical event time we specified 43 temporal profiles. Thanks to the named entity recognition module of Kyoto we are able to retrieve dates and, based on Wordnet, the system can recognize temporal expressions which refer to weekdays or months and more general and relative time markers (such as *now* or *two weeks later*).

Furthermore, 23 location profiles are utilized to extract geographical proper names and other locative expressions based on the Wordnet class of places (as *street, city, country* etc).

[14] For now we focused on human animate participants and those referred to by personal pronouns. In the future we will also look into extracting participants indirectly named through word combinations consisting of geo adjectives preceding words denoting weapons and transportation vehicles (such as *Serbian tanks*).

5 Evaluation

For the evaluation purposes we used the KYOTO triplet representation of historical events, which is a generic event representation format. A triplet consists of a historical action, mapped with its nearby occurring participant, location or time expression together with a label indicating the event slot type. In the evaluation the gold standard triplets will be compared with triplets generated by the system. A set of five texts from the Srebrenica corpus, written some years after the massacre, was tagged manually with historical events by two independent annotators. We obtained a very high inter-annotator agreement of 94% (0.91 Kappa).

As a baseline, we generated triplets from all constituent heads in a sentence. Each constituent head is once treated as an action while all the others are seen as participants. Applying the default relation – historical participant – the baseline achieved an average of 66% recall and a (understandably) low precision of less than 0.01%. Tables 1 and 2 present the performance of the system on the evaluation set. The abbreviations in the tables stand for: T. Nr – Token Number, G. Trp – Gold Triplets, S. Trp – System Triplets, C.S. Trp – Correct System Triplets, R – Recall, P. – Precision, F – F-measure.

Counts / File	T. Nr	G. Trp	S. Trp	C.S. Trp	R. %	P. %	F
File 1	243	5	4	1	20	25	0.22
File 2	440	32	25	18	56	72	0.63
File 3	647	58	68	32	55	47	0.51
File 4	429	32	22	17	53	77	0.63
File 5	209	19	19	12	63	63	0.63
Micro Average	-	-	-	-	49	57	0.53

Table 1. Evaluation results per file (micro average).

Counts / Relation	G. Trp	S. Trp	C.S. Trp	R. %	P. %	F
Participants	98	95	57	58	60	0.59
Time	17	20	13	76	65	0.70
Location	31	23	10	32	43	0.37

Table 2. Evaluation results per relation (macro average)

The system reached an overall recall of 49% and a precision of 57%. The low scores for file 1 can be explained by the fact that in this text some so called 'political events' were described such as responsibility issues and an investigation w. r. t. events in Srebrenica that was performed in the Netherlands few years after the massacre. Currently the system is not prepared to handle any other events than the conflict related ones.

Historical actions, evaluated in a separate non triplet evaluation cycle, were extracted with a recall of 67.94% and a precision of 51.96%. We extracted time expressions with the highest precision of 65% and also the highest recall of 76%. The lower recall and precision measures reached for the extraction of participants and especially locations can be explained by the type shift of the semantic class of locations used for referral to event participants. As mentioned before, so far we only are able to identify these if occurring in subject position; in the future we will add deeper syntactic dependency information into KAF and by that we will improve the recognition of locations used as participants.

6 Conclusion and Future Work

In this paper we showed that historical events can successfully be extracted from text, based on constructional clues and semantic type specification. To extract events we used a generic fact mining system KYOTO; we specified language structures and Wordnet concepts denoting event actions, participants, locations and time markers and we identified the historical events through recognition of historical actions. The evaluation results confirm that historical events can be extracted from historical texts by means of this approach with a relatively high recall of almost 50% and a precision of 57%, (comparable to the results of Wunderwald, 2011). In our future work we are going to increase the performance of the system by utilizing in the profiles more specific syntactic information and the grammatical tense. We will also look into other possibilities of distinguishing between historical events and events lacking historical value, also in non historical genres. In the next stage of the project we will make an attempt to automatically determine relations between historical events over textual data. We will also apply the system to other historical descriptions that are connected to museum collections. Because of the generic design of the extraction module, we expect that the extraction of conflict events can be applied to other periods and events with little adaptation.

Acknowledgments

This research was funded by the interfaculty research institute CAMeRA (Center for Advanced Media Research) of the VU University of Amsterdam: http://camera.vu.nl.

References

Agirre, Eneko 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, (EACL-2009), Athens, Greece.

Ahonen, Eeva and Eero Hyv¨onen, 2009, “Publishing Historical Texts on the Semantic Web -A Case Study” [online] available: http://www.seco.tkk.fi/publications/2009/ahonen-hyvonen-historical-texts-2009.pdf

Bosma, Wauter, Vossen, Piek, Soroa, Aitor, Rigau, German, Tesconi, Maurizio, Marchetti, Andrea, Monachini, Monica, and Carlo Aliprandi, 2009 “KAF: a generic semantic annotation format.”, in Proceedings of the GL2009 Workshop on Semantic Annotation, Pisa, Italy, Sept 17-19, 2009.

Bosma, Wauter and Piek Vossen, 2010, "Bootstrapping language neutral term extraction", in: Proceedings of the 7th international conference on Language Resources and Evaluation, (LREC2010), Valletta, Malta, May 17-23, 2010.

Cybulska, Agata and Piek Vossen, “Event models for Historical Perspectives: Determining Relations between High and Low Level Events in Text, Based on the Classification of Time, Location and Participants”, in Proceedings of LREC 2010, Valletta, Malta, May 17-23, 2010

Ide, Nancy and David Woolner, 2007, “Historical Ontologies”, in: Ahmad, Khurshid, Brewster, Christopher, and Mark Stevenson (eds.), Words and Intelligence II: Essays in Honor of Yorick Wilks, Springer, 137-152.

Tanev, Hristo, Piskorski, Jakub and Martin Atkinson, “Real-Time News Event Extraction for Global Crisis Monitoring”, in NLDB 2008: Kapetanios, Epaminondas, Sugumaran, Vijayan, Spiliopoulou, Myra (eds.) Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems, 2008, Springer: LNCS, vol. 5039, pp. 207-218.

Van Hage, Willem, Malaisé, Veronique, Segers, Roxane, Hollink, Laura (fc), Design and use of the Simple Event Model (SEM), the Journal of Web Semantics, Elsevier

Vossen, Piek, Agirre, Eneko, Calzolari, Nicoletta, Fellbaum, Christiane, Hsieh, Shu-kai, Huang, Chu-Ren, Isahara, Hitoshi, Kanzaki, Kyoko, Marchetti, Andrea, Monachini, Monica, Neri, Federico, Raffaelli, Remo, Rigau, German, Tescon, Maurizio, 2008a, "KYOTO: A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.

Vossen, Piek, Bosma, Wauter, Agirre, Eneko, Rigau, German and Aitor Soroa, 2010, "A full Knowledge Cycle for Semantic Interoperability", in: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010), Hong Kong, January 15-17, 2010.

Wunderwald, Martin, 2011, “NewsX Event Extraction from News Articles”, diploma thesis, Dresden University of Technology, Dresden, Germany, URL: http://www.rn.inf.tu-dresden.de/uploads/Studentische_Arbeiten/Diplomarbeit_Wunderwald_Martin.pdf

Xu, Feiyu, Uszkoreit, Hans, Li, Hong, 2006. “Automatic Event and Relation Detection with Seeds of Varying Complexity”, in: Proceedings of the AAAI 2006 Workshop Event Extraction and Synthesis, Boston, 491-498.