doi:10.5128/ERYa11.12

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 11, 189–203

EVENT COREFERENCE DETECTION IN ESTONIAN NEWS ARTICLES: PRELIMINARY EXPERIMENTS

Siim Orasmaa

Abstract. Event coreference detection is a task of automatically determining which fine-grained textual descriptions of events (e.g. sentences) corefer. The task is important in organizing information in large collections of news articles, as extracted coreferring event mentions can provide the user with an overview of the events discussed in articles, and can also provide a glimpse into the factual data related to the events. In this article, we survey previous approaches to automatic event analysis, discuss theoretical considerations related to event coreference detection and also outline a motivation for experimenting with event coreference detection in the context of limited linguistic resources. In the experimental part of our work, we consider the task of event coreference detection in the subset of articles mentioning a specific person in a fixed publishing period, and we use the experiments to outline possible general factors that are influencing the results.*

Keywords: natural language processing, text analysis, Estonian

1. Introduction

Event coreference detection is a task of automatically determining which textual event mentions corefer. In particular, the task is concerned with fine-grained text analysis: how to determine that two sentences or two words/phrases (two event mentions) are referring to the same event. For example, consider the following two sentences extracted from Estonian news:

(1) Soome sideminister Matti Aura andis eile hommikul lahkumispalve, sest tunnistas tehtud viga, kui ta hiljuti õigustas Vennamo käitumist.
'The Finnish Minister of Communications, Matti Aura, gave his resignation yesterday morning, as he admitted that his recent justification of Vennamo's behaviour was a mistake.' (ERC, Postimees 5.1.1999)

* This work was supported by Estonian Ministry of Education and Research (grant IUT 20-56 "Computational models for Estonian"), and by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS). The author also wishes to thank anonymous reviewers for helpful comments.

(2) Soome sideminister Matti Aura teatas eile oma tagasiastumisest seoses Sonera aktsiate müügi ümber puhkenud skandaaliga.
'The Finnish Minister of Communications, Matti Aura, announced his stepping down yesterday, related to the scandal of selling Sonera's shares.' (ERC, SL Õhtuleht 5.1.1999)

Both sentences (1) and (2) mention the same event – an announcement of resignation, which is described by two event mentions: 'gave his resignation' and 'announced his stepping down'.

From the perspective of a human reader, say, a historian trying to collect factual information regarding some events from contemporary newspapers, it is useful to compare different sources describing the same event in order to obtain a better picture of that event. This is because an event description from a single source can be incomplete (e.g. missing some factual information), or inaccurate (e.g. presenting wrong factual information). Factual information that is compatible across multiple/different sources can also be considered accurate with somewhat higher confidence (although the reliability of the source also plays an important role in determining the confidence, see Fokkens et al. 2014).

The task of automatically detecting coreferring event descriptions from news represents an interesting challenge in Natural Language Processing. As one can often assume that during the same publishing period (e.g. a day or a week), news articles from different sources tend to describe the same events, the search space of sentences potentially containing coreferring events can be effectively narrowed down by time constraints. Within such constrained search space, even a simple method matching lemmas of event mentions could be sufficient for solving the task (Cybulska, Vossen 2014). Note that while many past attempts at sentence-level event analysis have relied on conceptual modelling of events (e.g. relying on a set of predefined semantic frames for event detection), event coreference detection allows one to explore more simple and general event models, such as models where an event is mainly identified by arguments describing its participants, time and location (Glavaš, Šnajder 2013, Cybulska, Vossen 2013). Still, the task is very difficult, and difficulties seem to stem from the abstractness of the notion of event. It is not clear, how the notion of event should be established linguistically: which size text units serve best as concise descriptions of events (words/phrases/clauses/sentences/...?). Furthermore, even if the notion of event is established at some level, the coreference between events can still be hard to decide, as one needs to consider other relations between events, such as subevent and membership relations (Hovy et al. 2013).

The present work explores event coreference detection in Estonian news articles. We consider a realistic information retrieval setting, where the user wants to find information about events related to a specific person from a corpus of daily newspapers. In the experiments, we test out simple heuristics (lemma matching, as well as matching of event arguments of location and time) for the task of finding pairs of sentences mentioning the same event. Our work is preliminary as we do not aim to provide a thoroughly evaluated method for solving the problem. Instead, the aim is to explore how general factors – lexical homogeneity within the set of articles under analysis, and the media coverage of the events related to the person – influence the results.

2. A few notes on the notion of event

It is difficult to give a good definition for the notion of **event**, as many definitions seem to involve synonymous abstract words, such as *occurrence* or *happening*, and using these words can ultimately lead to a circular definition. The question about the essence of all events has been studied in philosophy, and there it eventually led to the question of event identity: when can two events be considered identical (Schneider 2005)? Donald Davidson (1969) first proposed that two events having the same causes and effects should be considered as identical. Later he rejected the account and agreed with Willard Van Orman Quine (1985) that two events occurring at the same time and in the same space should be considered identical (Davidson 1985).

In this work, we try to follow the Quinean-Davidsonian notion of event (an event being identified by its spatiotemporal location) in making judgments concerning coreference relations between events mentioned in natural language texts. As this is a preliminary work, we make no further attempt to divide events into classes, and we also regard states (such as *Jaan is hungry*, the *oven is hot*) as events.

3. Different levels of event analysis in information retrieval and extraction

As the concept of event is an abstract one, it allows a number of different views on how events are expressed in natural language. In this work, we focus mainly on information retrieval and extraction context, where one has a practical goal of finding documents (news articles) talking about the same event(s), and extracting fine-grained details about the events of interest. Previous work in this field can be roughly divided into three event analysis levels, each considering a different size text unit as a description of an event: **document-level**, **sentence-level**, and **mention-level** analysis.[1]

Document-level event analysis mostly relies on an assumption that an article/news story has been built around one main event (the topic), although it also describes relevant events related to the main event. In Topic Detection and Tracking research (Allen et al. 1998) the goal is to analyze a stream of news stories and to detect stories that are discussing the same or directly related events. Because an event has a rather large textual description in this view (the whole article), lexical similarity between articles can be relatively successfully utilized for solving the task. A limitation of this processing level is that articles are often discussing many events, and knowing that two articles are discussing the same event does not immediately tell us which event exactly is the common one, nor does it reveal any interesting details of the event (e.g. who did what to whom, when and where?).

Sentence-level event analysis focuses on detecting an event mention along with a detailed description of its structure: event participants and circumstances. It is, perhaps, inspired by a prototypical simple sentence, in which the main verb refers to the event and other complements refer to participants and circumstances of the event (event arguments). On this level of event analysis, it is often assumed that

[1] This division draws inspiration from a presentation by TimeML Working Group (2007).

the conceptual structures of events are predefined, e.g. in the form of FrameNet semantic frames (Baker et al. 1998) or ACE[2] event types, and the goal is to detect presence of the event in the sentence, and to fill the argument slots in the corresponding conceptual structure. However, matching conceptual semantic arguments with their actual realizations in text has proven to be a very difficult task, and so far, successful automatic event analysis on this level has required setting the focus on a narrow set of event types (e.g. only events of *attack*, *death* and *injury*) (Naughton 2009). The obvious limitation is that many other events that are mentioned in the text will be left unanalysed (including relevant events not well-described in the given set of event types).

Mention-level event analysis focuses on the smallest possible linguistic units describing an event – event mentions such as verbs (e.g. *married*), nouns (e.g. *wedding*) and adjectives (e.g. *(was) pregnant*). Because the focus is only on detection of event mentions, without accompanying event argument structures, this makes it possible to define a rather general event model, which has the advantage of not being restricted to a list of predefined event types (like FrameNet frames or ACE event types). The disadvantage is, however, that it is difficult to give a precise definition of which linguistic units should be considered events. Complex problems include analysing multiword event units (e.g. does *held a conference* indicate one or two events, as the act of holding/organizing a conference can be distinguished from the actual event – the conference – as the former usually takes a longer time span), and distinguishing between event and non-event readings of nominals (e.g. *dinner* as an event and as a food) (Sprugnoli, Lenci 2014). TimeML (Pustejovsky et al. 2003) is an example of a framework that focuses on event annotation at the mention level.

4. Event coreference detection

4.1. Theoretical and empirical concerns regarding the event coreference detection

The task of event coreference detection can be divided into **within-document event coreference detection** (i.e. finding coreferring event mentions within the same document) and **cross-document event coreference detection** (i.e. finding coreferring event mentions from different documents) (Naughton 2009).

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki and Andrew Philpot (2013) note that the difference between coreference and non-coreference is not clear-cut, and partial coreference of events can be distinguished. They study partial coreference in detail, arguing that partial coreference could indicate either a membership or a subevent relation between two events. A membership relation holds when one event mention describes a set of events of the same type, and the other event mention refers to a single member of that set. For example, in the sentence "I had *three meetings* last week, but only *the last one* was constructive" the event referred to as *the last one* can be considered a member of the set of events *three meetings*. In the case of a subevent relation, one event is considered to be a stereotypical sequence of events having a common goal (a script, such as *eating at the restaurant*, *visiting a doctor*), and the other event (the subevent) is one of

[2] ACE (Automatic Content Extraction) programme – a programme that was established in United States for developing automatic Information Extraction technologies, see also https://www.ldc.upenn.edu/collaborations/past-projects/ace (4.9.2014)

the actions/events executed as part of that script (e.g. *ordering the meal* is part of *eating at the restaurant*, and *making an appointment with the doctor* is part of *visiting a doctor*). Authors also argue that the time, location, and participants of events provide key information for distinguishing between different types of coreference (and non-coreference). However, their inter-annotator agreement results on distinguishing partial coreference relations indicate that these concepts are not yet well understood.

The idea that time, location, and participants could play an essential role in event coreference detection has also been explored by Goren Glavaš and Jan Šnajder (2013). The authors use a generic event model where an event is defined by an event anchor (an event mention, usually a single word), and one or more arguments of four broad semantic types (agent, target, time, and location). They note and also confirm by inter-annotator agreement experiments that when humans are asked to decide event coreference considering only mentions and four arguments, their agreement and certainty on the task is rather low, as opposed to when they can examine the full context (both documents where event mentions occurred) before making the decision. This suggests that event coreference detection is inherently a composite task, where all levels of event analysis – the document, sentence, and mention levels – should be considered together.

4.2. Event coreference detection in the context of limited linguistic resources

Most recent works on automatic event coreference detection have considered English or other languages well-equipped with linguistic resources, such as Spanish, Italian and Dutch (Glavaš, Šnajder 2013, Cybulska, Vossen 2013, Vossen et al. 2014). The task in its full complexity seems to require that a number of advanced language analysis steps are implemented: 1) detection of event mentions, 2) named entity recognition and semantic role labelling for detection of event arguments, 3) normalization of temporal and locational expressions[3] and named entity coreference resolution for aligning event argument structures, 4) aligning event mentions along with their argument structures. However, it is not clear to what extent these fine-grained language analysis steps must be (and can be) solved, and some of the recent research suggests that light-weight approaches to event coreference detection are also worth trying out.

Firstly, it is likely that tasks of within-document event coreference and cross-document event coreference have different levels of difficulty. Martina Naughton (2009) notes that two sentences mentioning the same event within the same document are likely to have heterogenous vocabulary, as the factual information (e.g. location, participants) is rarely repeated on the second mention of the event. In contrast, two cross-document sentences mentioning the same event have likely more homogeneous vocabulary, because the important factual information (location, time, participants) is repeated in both documents. Thus, the task of cross-document event coreference can potentially be approached even with simple methods relying on lexical similarity.

[3] E.g. finding calendar dates corresponding to temporal expressions, and finding coordinates of geographical regions corresponding to locational expressions.

Secondly, if the set of documents under analysis can be narrowed down to the subset of documents discussing the same event(s) (such as in the Topic Detection and Tracking task), even a simple method – matching event mentions by their lemmas – can yield relatively good results, which are roughly comparable to the results obtained with complex methods matching event argument structures. This suggests that document clustering plays an important role in event coreference detection, and if one can obtain clusters of documents discussing the same events (e.g. by using lexical similarity and time constraints), one can resolve the amount of mention-level event coreference within these clusters by using simple lexical similarity methods. (Cybulska, Vossen 2014)

In this work, we explore the problem of event coreference detection in Estonian. As this is a preliminary work on the task in Estonian, we also start out experimenting with simple methods.

5. A case study: finding person-related events from news articles

As a case study, we consider an exploratory search setting, where the user wants to find events related to a specific person from a corpus of daily newspapers. We make a crude assumption that such events can be found from articles where the person is mentioned by full name, and the goal is to find pairs of sentences mentioning the same event within this set of articles.[4] We add two additional constraints. Firstly, we only consider articles from a fixed time period: one week. And secondly, we focus on finding pairs of sentences from different articles, i.e. on cross-document event coreference.

5.1. Document-level considerations regarding event coreference

Before we can begin exploring sentence-level event coreference, we likely want to have some rough characterization of the given set of articles regarding the possible difficulty of sentence-level event coreference. Lexical similarity within the set of articles (more specifically, the average lexical similarity between all pairs of articles within the set of articles) can serve as such a characteristic:

- Lexically homogeneous articles (with high lexical similarity) are likely describing the same events.
- Lexically heterogeneous articles (with low lexical similarity) can describe different events, or can describe the same events using rather different vocabularies.

Naturally, the lexical heterogeneity is affected by the timespan chosen for analysis: the longer timespan typically means that there are more events reported, thus a more lexically heterogeneous set of articles is obtained. However, as we consider only articles mentioning a specific person, the increase of reports also depends on how much the media covers the given person's activities, and how this coverage changes over time.

[4] An alternative would be to attempt to detect clusters of sentences, so that each cluster contains sentences referring to same event. However, this approach would be more difficult to evaluate (considering partial coreference relations and the fact that one sentence can refer to multiple events), so we chose to explore the pair-wise detection approach instead.

In the present work, we use the vector space model and calculate cosine similarity over the word lemmas to measure lexical similarity between articles. We also apply tf-idf weighting for lemmas (Manning et al. 2008), which means that lemmas occurring many times in few articles have the highest discriminative power (and sharing such lemmas contributes the most to the similarity between articles), while lemmas occurring almost in all articles of the corpus have the lowest discriminative power (thus sharing such lemmas contributes only little to the similarity between documents). Lemma counts and occurrences in articles are calculated over the whole corpus (one week of news articles).

5.2. Sentence-level event coreference

In our experimental setup (details discussed in the next section), we consider 3 different sets of articles with varying degrees of lexical similarity, and we test two different methods for sentence-level event coreference detection: 1) a simple lexical similarity measure (measuring the amount of lemmas overlapping between two sentences), and 2) an overlap of event argument structure components (location and time). It is expected that if the set of articles is lexically homogeneous, a simple lexical similarity measure (method 1) yields high-precision results, which potentially outperforms (in terms of precision) a more sophisticated event model involving argument structure similarity (method 2). On the other hand, in a lexically heterogeneous corpus, the simple lexical similarity measure is likely ineffective, and the model matching event location and time will likely provide better results.

As a similarity measure, we use the Jaccard Similarity Coefficient (Jaccard 1901), which is defined as a similarity between two sets: the size of the intersection of two sets divided by the size of the union of two sets.

In order to find the simple lexical similarity between two sentences (method 1), we extract word lemmas from both sentences, convert to sets (i.e. remove duplicate lemmas) and calculate the Jaccard Similarity Coefficient between these sets of lemmas. All pairs of sentences that had a similarity coefficient value greater than or equal to 0.5 were selected as pairs potentially containing coreferring event mentions.

Note that one of the possible limitations of using the Jaccard Similarity Coefficient over whole sentences is that the measure is sensitive to the contrast between sentence lengths: if a short sentence entirely overlaps with a long one, and the short one is less than half the length of the longer one, a coefficient value below 0.5 is obtained. This problem can be alleviated by using a method less sensitive to the difference between sentence lengths: the Second Kulczynski Coefficient (Pecina 2010). This method finds an average of two ratios: the size of the intersection divided by the cardinality of the first set, and the size of the intersection divided by the cardinality of the second set. However, as our experiments showed, this measure introduces another problem: a short sentence containing mainly non-content words (such as “Aga kes siis veel” ‘But who else then’) can match with many long sentences containing these non-content words, thus introducing some degree of noise into the results. Although filtering of non-content words could potentially alleviate this problem, we chose to stay with the Jaccard Similarity Coefficient in order to keep our models simple.

In the second similarity method, we consider overlap of argument structure components of the event mentions: temporal expressions and location names mentioned in sentences. Because we don't have syntactic structure or semantic role annotations available, we use a very crude approximation and consider all temporal expressions and locations appearing in one sentence to belong to one event argument structure. For temporal expressions, we consider normalized calendrical values of the expressions instead of lemmas, which allows us to match (lexically) different date expressions, such as *täna* 'today' and *eelmisel neljapäeval* 'last Thursday' if they refer to the same normalized date (e.g. 11.9.2014; for more on annotation of temporal expressions in Estonian, see Orasmaa 2012). We focus on date expressions containing day granularity (e.g. *two days ago*), month granularity (e.g. *in April*), or year granularity (e.g. *last year*) temporal information. For location names, we only consider lemma matches as we currently do not have means for normalizing location expressions to coordinates of geographical regions.

Similarly to the first method, we apply the Jaccard Similarity Coefficient and calculate two scores: a coefficient for matching calendrical values of temporal expressions, and a coefficient for matching lemmas of location expressions. If both coefficient values were greater than or equal to 0.5, then the pair of sentences was considered as potentially containing coreferring event mentions.

6. Experiments

6.1. The corpus

In our experiment, we selected all news articles from one week (from 4.1.1999 to 10.01.1999) of three Estonian daily newspapers (Postimees, Eesti Päevaleht, and SL Õhtuleht), as they can be found in the Estonian Reference Corpus[5] (Kaalep et al. 2010). The corpus has been automatically annotated for sentence boundaries and morphological information (word lemmas, part of speech tags, morphological case and conjugation information). In addition, there are two layers of automatically added factual/semantic annotations: named entities (persons, organizations, locations, addresses, quantities), and temporal expressions, and the latter are normalized according to the TimeML annotation format (Pustejovsky et al. 2003).

For our experiment, we have chosen three persons – Jüri Mosin, Pekka Vennamo and Mart Siimann, as articles mentioning these persons are characterized by different lexical similarity levels (measured as an average of cosine similarities between all pairs of articles), and all of these article sets contain articles from three different newspapers. Table 1 reports the statistics related to the three article sets.

[5] http://www.cl.ut.ee/korpused/segakorpus/ (7.2.2015).

Table 1. Statistics of article sets mentioning the three given persons

Person	Articles	Sentences	Words	Avg cosinus similarity (tf-idf weighting)
jüri Mosin	4	64	1080	0.56
Pekka Vennamo	9	142	2610	0.45
Mart Siimann	27	624	10507	0.13

We will briefly describe the events mentioned in given article sets. Articles mentioning Jüri Mosin were mostly focusing on a single event – a trial over criminal Jüri Mosin and his accomplices – although they also discussed background events, such as the criminal history of the persons under trial, and possible future appeals. In terms of events, the articles were rather focused, which is also reflected in high lexical homogeneity. Articles mentioning Pekka Vennamo were discussing a scandal related to the person: a possible abuse of his official position for purchasing shares. The scandal had a longer development history, and on the given week, it culminated in the firing of Pekka Vennamo from his position, and the resignation of a Finnish minister related to the scandal. In terms of events, the articles were still quite focused on the two events (the firing and resignation of two high officials), although the history of the scandal, possible future developments and effects were more widely discussed than in the first case. Articles mentioning Mart Siimann, who was the prime minister of Estonia at that time, were least focused on certain events. There were few stories where his actions were at the focus, such as meetings with prime ministers of neighbouring countries; however, many articles just mentioned him once or twice, in a discussion of events not directly related to him (e.g. an article describing a sports event mentioned once that the prime minister was supporting the event). This lack of focus on a narrow set of events is also reflected in low lexical similarity between articles.

6.2. Evaluation of sentence-level event coreference

Because we do not have a corpus annotated for event coreference available, we can only measure and compare methods in terms of precision (that is, the number of correct sentence pairs divided by the number of all pairs of sentences returned by the method).[6]

Sentence pairs found by both methods were evaluated manually. If two sentences under comparison contained multiple event mentions, it was required that only one pair of event mentions should corefer for the sentence pair to be correct. For example, consider the following pair of sentences:

(3) Läti peaminister Vilis Krishtopans lubas eile kohtumisel Eesti valitsusjuhi Mart Siimanniga ühtlustada kahe riigi vahelist viisarezhiimi. (ERC, Eesti Päevaleht 8.1.1999)
'Latvian prime minister Vilis Krishtopans promised to homogenise the visa regime between the two countries at yesterday's meeting with Estonian prime minister Mart Siimann.'

[6] A limitation of measuring only precision is that we do not have an overview of the completeness of the results (which can be measured by recall). However, measuring recall requires creating a corpus annotated for event coreference, which is not a trivial task (due to the complexity of the phenomenon). We believe that it requires a separate study, which is outside the scope of this preliminary work.

(4) Eesti-Läti sealihatüli oli peaminister Mart Siimanni ja tema Läti kolleegi, peaminister Vilis Krishtopansi eilse kohtumise üks põhiteemasid. (ERC, Postimees, 8.1.1999)
'The Estonian-Latvian dispute over pork trade was one of the main subjects of yesterday's meeting between prime minister Mart Siimann and his Latvian colleague Vilis Krishtopans.'

Despite the fact that event mentions such as *promised, to homogenise*, and *dispute* are not common between sentences (3) and (4), we considered this pair to be correct because the *meeting* (of two prime ministers) is mentioned in both sentences.

The cases of uncertainty on deciding event coreference, which in our case also included the partial coreference discussed by Hovy et al. (2013), were counted separately from correct/incorrect cases.

Results of manual evaluation of the simple lexical similarity method (method 1) are listed in Table 2. As expected, the method obtains high precision on lexically homogeneous sets of articles (Jüri Mosin, Pekka Vennamo), and obtains more problematic results on the lexically most heterogeneous set of articles (Mart Siimann). However, the total number of found pairs is too small to draw any strong conclusions.

Table 2. Results of similarity method 1 (Jaccard Similarity Coefficient for measuring lemma overlap between two sentences) on finding sentence pairs referring to the same event

Person	Correct pairs	Incorrect pairs	Pairs with uncertain coreference	Precision
Jüri Mosin	4	0	0	100%
Pekka Vennamo	12	0	0	100%
Mart Siimann	4	0	2	66.6%
Total	20	0	2	90.9%

The results of manual evaluation of method 2 (overlap of calendric values of temporal expressions and lemmas of location expression) are listed in Table 3. Similarly to method 1, this method returns high-precision results on the lexically most homogeneous set of articles (Jüri Mosin), but it does not outperform method 1 on the lexically more heterogeneous sets of articles (Pekka Vennamo, Mart Siimann) as was expected. However, it is notable that method 2 makes rather different kinds of mistakes on these datasets – on the Pekka Vennamo dataset it returns more pairs with uncertain coreference, while in the Mart Siimann dataset it returns more pairs with incorrect coreference.

Table 3. Results of similarity method 2 (Jaccard Similarity Coefficient for measuring overlap of temporal and locational expressions of two sentences) on finding sentence pairs referring to the same event

Person	Correct pairs	Incorrect pairs	Pairs with uncertain coreference	Precision
Jüri Mosin	3	0	0	100%
Pekka Vennamo	6	2	5	46.1%
Mart Siimann	11	10	2	47.8%
Total	20	12	7	51.2%

We also examined the degree of overlap between the results returned by the two methods. It turned out that only 2 sentence pairs (out of 61 pairs returned by the two methods together) were overlapping, which shows the potential for these two methods to complement each other.

7. Discussion and conclusion

The dataset used in our evaluation is too small to draw any strong conclusions, but the general impression is that the results of fine-grained event coreference detection are dependent on how focused the articles are on a narrow set of events. This narrow set of events is difficult to describe formally, but it seems to manifest itself in a way that it is possible to name central/main events covered by the media in a given time period: in the case of Jüri Mosin, the main event was the trial, and in the case of Pekka Vennamo, the main event was the scandal. On the other hand, it is difficult to name one or even a few central events in the set of articles mentioning Mart Siimann, as he was mentioned in relation to a rather diverse set of events.

Naturally, if the set of articles is focused on a narrow set of events, there is less room for making mistakes, and even a minimal lexical similarity can indicate that both sentences are likely mentioning the same events. And if the set of articles is not focused, but describes a large set of events, it is likely that even more sophisticated coreference detection methods (compared to the simple lexical similarity) will be fallible, as there can be misleading similarities between event descriptions actually referring to different events.

We hypothesize that the media coverage of the events related to the person plays an important, albeit a latent role in narrowing down the set of events. Our dataset indicates that if the person is only mentioned in a few articles, the stories are likely built around a narrow set of events, while frequent mentions more likely indicate a large set of events. As discussed by Johan Galtung and Mari Holmboe Ruge (1965), various factors can explain these tendencies. One can note that events related to elite persons are reported more often, and events related to non-elite persons are reported infrequently and more likely if they are negative (factors that both contribute to the reports being more focused). Also, if the person has high media coverage, she/he is more likely mentioned in opinions and discussions, which are less concerned with reporting factual details about events, while low media coverage could indicate that the reports are more factual in nature. Considering this, we can speculate that it is easier to find coreferring event mentions if a person has low media coverage, as there is less room for mistakes (only events reported inside the fixed time period are likely the coreferring ones) and potentially more factual information.

Returning to our dataset and to the problem of whether the article set is focused on a narrow set of events, high lexical similarity seems to be one indicator of a narrow set of events. The sets of articles mentioning Jüri Mosin and Pekka Vennamo were both characterized by relatively high lexical similarity, compared to the set of articles mentioning Mart Siimann (Table 1). However, this indicator is affected by the length of the timeframe chosen for analysis, and one cannot rule out the possibility that two articles having relatively low lexical similarity are still discussing the same events, using different vocabularies.

In terms of fine-grained event coreference detection and evaluation, our results indicate that the simple lexical similarity method (lemma overlap between sentences) seems to be the most stable method for producing high-precision results. However, it is also the easiest method to judge: high lexical overlap between words leaves little room for doubt on whether two sentences refer to the same events or not. When we loosen the lexical similarity constraints and allow coreferring events to be detected only by words indicating the spatial and temporal context of events (the method 2), the task of judging event coreference becomes more difficult, and there will be more cases of uncertainty. The uncertainty indicates that there is some structural similarity between events mentioned in two sentences, so we cannot decide for sure that no coreference exists. Consider an example of an uncertain match from the set of articles mentioning Pekka Vennamo (returned by method 2):

(5) Vennamo ise ütles eile Soome telekanalile MTV 3, et ei teinud midagi valesti ega ole andnud ka valeinfot Sonera aktsiatega sooritatud tehingute kohta. (ERC, Eesti Päevaleht 5.1.1999)
'Vennamo himself told the Finnish TV channel MTV 3 yesterday that he had not done anything wrong and he had not given any wrong information regarding the transactions with Sonera's shares.'

(6) Varem Vennamot toetanud Soome sideminister Matti Aura loobus skandaali tõttu teda usaldamast ja astus eile ka ise tagasi. (ERC, Postimees 5.1.1999)
'A former supporter of Vennamo, Finnish Minister of Communications Matti Aura, stopped trusting Vennamo because of the scandal and also resigned yesterday from his position.'

The main event mentions of the sentences (i.e. 'told', 'had not done' and 'had not given' in sentence (5), and 'stopped' and 'resigned' in sentence (6)) are referring to events performed by different persons and thus cannot corefer. However, the event mention 'scandal' in sentence (6) refers to a rather general event (the event which was also the main reason why both mentioned individuals ended up in the spotlight of Estonian media in the given week), and arguably one can consider the events 'told', 'stopped trusting' and 'resignation' all as subevents of the event 'scandal', thus indicating a partial coreference according to Hovy et al. (2013).

Note that the uncertainty introduced when we loosen the lexical similarity constraints (moving from method 1 to method 2) can also be used as an indicator of the degree to which the set of articles is focused on a narrow set of events. In the set of articles mentioning Pekka Vennamo, switching from method 1 to method 2 introduces more uncertain pairs than incorrect pairs, while in the set of articles mentioning Mart Siimann, the switching introduces more incorrect pairs than uncertain pairs. If there are more uncertain than incorrect pairs, this indicates that the set of events is probably narrow, but the descriptions are revealing the structure of events (e.g. subevents and member events of the focused events are being described), so it is difficult to decide whether we have a coreference or not. However, if loosening the lexical similarity constraints introduces more incorrect than uncertain pairs, the set of articles is likely not focused enough on a narrow set of events.

In conclusion, our preliminary results confirm the findings of previous works that relatively high lexical similarity within a (sub)set of articles can provide an important

indicator that the set of articles is focused on a narrow set of events, and high-precision event coreference detection can be performed on the dataset. However, our main contribution is to outline the importance of media coverage of the events (related to the person searched for) as an underlying mechanism influencing the results. An interesting question is whether we can somehow formalize this notion and its characteristics influencing how factually (or non-factually) the events are discussed in media – which events or persons tend to evoke more factual reports. To some extent, the presence of the factors influencing the media coverage discussed by Galtung and Ruge (1965) can be detected, e.g. one can determine the negativity of the text by using automatic sentiment analysis.

References

Allan, James; Carbonell, Jaime G.; Doddington, George; Yamron, Jonathan; Yang, Yiming 1998. Topic detection and tracking pilot study final report. – Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, February 8–11, 1998. http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa98/pdf/tdt2040.pdf (28.2.2015).

Baker, Collin F.; Fillmore, Charles J.; Lowe, John B. 1998. The Berkeley Framenet Project. –Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. 1, ACL '98. Stroudsburg, PA: Association for Computational Linguistics, 86–90. http://dx.doi.org/10.3115/980845.980860

Cybulska, Agata; Vossen, Piek 2013. Semantic relations between events and their time, locations and participants for event coreference resolution. – Recent Advances in Natural Language Processing. Proceedings, 156–163.

Cybulska, Agata; Vossen, Piek 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 4545–4552.

Davidson, Donald 1969. The individuation of events. – N. Rescher et al. (Eds.). Essays in Honor of Carl G. Hempel. Synthese Library 24. Dordrecht: Reidel, 216–234. Reprinted in D. Davidson, Essays on Actions and Events. Oxford: Clarendon Press, 2001, 163–180. http://dx.doi.org/10.1093/0199246270.003.0008

Davidson, Donald 1985. Reply to Quine on events. – E. LePore, B. McLaughlin (Eds.). Actions and Events: Perspectives on the Philosophy of Donald Davidson. Oxford–New York: Basil Blackwell, 172–176.

Fokkens, Antske; Braake, Serge ter; Ockeloen, Niels; Vossen, Piek; Legêne, Susan; Schreiber, Guus 2014. BiographyNet: Methodological issues when NLP supports historical research. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3728–3735.

Galtung, Johan; Ruge, Mari Holmboe 1965. The structure of foreign news the presentation of the Congo, Cuba and Cyprus Crises in four Norwegian newspapers. – Journal of Peace Research, 2 (1), 64–90. http://dx.doi.org/10.1177/002234336500200104

Glavaš, Goran; Šnajder, Jan 2013. Exploring coreference uncertainty of generically extracted event mentions. – Alexander Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. 14th International Conference, CICLing 2013, Samos, Greece, March 24–30, 2013, Proceedings, Part 1. Berlin–Heidelberg: Springer, 408–422. http://dx.doi.org/10.1007/978-3-642-37247-6_33

Hovy, Eduard; Mitamura, Teruko; Verdejo, Felisa; Araki, Jun; Philpot, Andrew 2013. Events are not simple: Identity, non-identity, and quasi-identity. – Proceedings of the

1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Atlanta, Georgia, 14 June 2013. Association for Computational Linguistics, 21–28.
Jaccard, Paul 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. – Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547–579.
Kaalep, Heiki-Jaan; Muischnek, Kadri; Uiboaed, Kristel; Veskis, Kaarel 2010. The Estonian Reference Corpus: Its composition and morphology-aware user interface. – I. Skadina, A. Vasiljevs (Ed.). Human Language Technologies – The Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT. Frontiers in Artificial Intelligence and Applications 219. IOS Press, 143–146.
Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich 2008. Introduction to Information Retrieval. Vol. 1. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809071
Naughton, Martina 2009. Sentence Level Event Detection and Coreference Resolution. PhD dissertation. Dublin: National University of Ireland.
Orasmaa, Siim 2012. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. [Automatic recognition and normalization of temporal expressions in Estonian language texts.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 153–169. http://dx.doi.org/10.5128/ERYa8.10
Pecina, Pavel 2010. Lexical association measures and collocation extraction. – Language Resources and Evaluation, 44 (1–2), 137–158. http://dx.doi.org/10.1007/s10579-009-9101-4
Pustejovsky, James; Castaño, José M.; Ingria, Robert; Sauri, Roser; Gaizauskas, Robert J.; Setzer, Andrea; Katz, Graham; Radev, Dragomir R. 2003. TimeML: Robust specification of event and temporal expressions in text. – New Directions in Question Answering, 3, 28–34.
Quine, Willard Van Orman 1985. Events and reification. – E. LePore, B. P. McLaughlin (Eds.). Actions and Events: Perspectives on the Philosophy of Donald Davidson. Oxford–New York: Basil Blackwell, 162–171.
Schneider, Susan 2005. Internet Encyclopedia of Philosophy: Events. http://www.iep.utm.edu/events/ (7.9.2014).
Sprugnoli, Rachele; Lenci, Alessandro 2014. Crowdsourcing for the identification of event nominals: An experiment. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 1949–1955.
TimeML Working Group 2007. Semantic Annotation: A TimeML Case Study. http://semanticweb.kaist.ac.kr/research/tc37sc4/new_doc/ISO_TC37_SC4_N337_WG2_ISO-TimeML_Tilburg2007.pdf (9.2.2015).
Vossen, Piek; Rigau, German; Serafini, Luciano; Stouten, Pim; Irving, Francis; Hage, Willem Van 2014. NewsReader: Recording history from daily news streams. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2000–2007.

Web references

ACE. https://www.ldc.upenn.edu/collaborations/past-projects/ace (4.9.2014).
Estonian Reference Corpus. http://www.cl.ut.ee/korpused/segakorpus/ (7.2.2015).

Siim Orasmaa (University of Tartu). Research interests: topics related to information extraction, with the specific focus on extraction of event and temporal information from natural language texts.
J. Liivi Str 2, 50409 Tartu, Estonia
siim.orasmaa@ut.ee

SÜNDMUST VÄLJENDAVATE KEELENDITE SAMAVIITELISUSE TUVASTAMINE EESTI KEELE UUDISTEKSTIDES: ESIALGSED KATSETUSED

Siim Orasmaa
Tartu Ülikool

Artiklis uuritakse, kuidas automaatselt tuvastada samal ajaperioodil ilmunud ajaleheartiklitest samadele sündmustele viitavaid lauseid. Uurimisküsimusele keskendutakse eeskätt infootsingute kontekstis, kus samadele sündmustele viitavad laused peaksid kasutajale pakkuma esmast ülevaadet sündmuste meediakajastusest (st võimaldama erinevate viitamiskontekstide ja allikate kõrvutamist ning faktilise info esmast kontrollimist).

Antakse ülevaade lähenemisviisidest, mida senised tööd on uudistekstide sündmusanalüüsil kasutanud: eristatakse sündmusanalüüsi dokumendi, lause ja sõna/fraasi tasemel. Seejärel tutvustatakse probleemi teoreetilisi aluseid (osalise ja täieliku samaviitelisuse eristamist) ning tuuakse välja võimalikud lähte-eeldused (artiklitevaheline samaviitelisuse tuvastamine on lihtsam kui artiklisisene; kõrge artiklitevaheline leksikaalne sarnasus võimaldab ennustada sündmuste kattuvust), millele tuginedes võib katsetada ka lihtsamaid (peamiselt leksikaalsele sarnasusele tuginevaid) meetodeid samaviitelisuse tuvastamiseks.

Vaadeldakse infootsingu ülesannet, kus kolme päevalehe ühe nädala artiklite koguhulgast tuleb leida ühe konkreetse isikuga seotud sündmusi. Samale sündmusele viitavaid lauseid otsiti kõigist artiklitest, kus isikut mainiti nimepidi. Katsetati kahte Jaccard'i koefitsiendil põhinevat meetodit samaviitelisuse tuvastamiseks: lause sõnalemmade kokkulangevusel põhinev meetod ning aja- ja kohaväljendite kokkulangevusel põhinev meetod. Eksperimendis valiti välja kolm isikut, kellega seotud artiklite hulki iseloomustab erinev leksikaalse sarnasuse tase, ning uuriti meetodite täpsust nendel artiklihulkadel. Esialgsed tulemused kinnitasid varasemaid tähelepanekuid, et suure artiklitevahelise leksikaalse sarnasuse kontekstis võivad ka lihtsad meetodid anda kõrge täpsusega tulemusi. Samas paistab oluliseks tulemust mõjutavaks teguriks olevat just isikuga seotud meediakajastuse iseloom: kitsale sündmuste hulgale fokuseeritud kajastuste korral on ülesanne ilmselt lihtsam kui juhtudel, mil isikut mainitakse paljude erinevate sündmuste kontekstis. Mil määral meediakajastuse iseloomu saab formaliseerida ning rakendada sündmust väljendavate keelendite samaviitelisuse tuvastamisel (tuvastamise kvaliteedi ennustamisel) ning kuivõrd siinsete esialgsete katsete tulemused leiavad kinnitust ka suuremate andmehulkade analüüsil, jääb aga uurimiseks edaspidi.

Võtmesõnad: automaatne keeletöötlus, tekstianalüüs, eesti keel