

Extraction of Temporal Expressions from Finnish News-feed

Juha Makkonen and Helena Ahonen-Myka
Department of Computer Science
University of Helsinki
P.O. Box 26, 00014 University of Helsinki, Finland
{jamakkon,hahonen}@cs.helsinki.fi

Abstract

The harnessing of time-related information from text for the use of event detection, for example, requires a leap from the surface forms of the expressions to a formalized time-axis. We present a methodology for extraction of Finnish temporal expressions and a scheme of comparing the temporal evidence of the news documents. We employ the comparison in identifying news events.

1 Introduction

News documents contain a wealth of information coded in the natural language temporal expressions. Automatic processing of news often neglects these expressions for several reasons: temporal expressions are difficult to spot, their surface forms themselves are hardly of any direct use, the meaning of an expression is often somewhat ambiguous or at least very difficult to resolve, and the expressions do not lend themselves easily to any kind of comparison. However, this temporal information would be highly useful for many areas of applications: (context aware) information retrieval, information extraction, question answering, document summarization, and topic detection and tracking, for example (Setzer, 2001).

There are three problems one has to deal with before temporal information can be applied automatically in any of these tasks: recognition, formalization and comparison of the temporal expressions. First, the expressions have to be extracted from the text. Second, the expressions need to be provided with formal meaning. In addition to the expression, one often needs some context information, such as the utterance time and the tense of the relevant verb, in order to map the expression. And finally, there has to be a suitable method of employing the formalized expression.

In this paper, we present an extraction and formalization approach for Finnish temporal expressions. Furthermore, we examine temporal similarity of news documents in the context of topic detection and tracking (TDT) (Allan et al., 1998; Allan, 2002) where one attempts to detect new, previously unreported events from online news-stream and track documents discussing the same event. Clearly, an event as well as the news-stream itself are intrinsically sensitive to time. In TDT, an event is usually understood as "some unique thing that happens at some point in time". This definition differs from the concept of event in artificial intelligence and dialogue systems, for example.

In recognizing temporal expressions, we employ syntactical parsing and finite-state automatas. Once an expression is recognized, the terms it contains are converted to shift and span operations that move the utterance time to the past or to the future. We define these operations on top of a calendar that defines the global timeline and its various granularities of time (Goralwalla et al., 2001). Each expression is formalized as a pair of dates on a global timeline.

Finally, we outline an approach to comparing the temporal similarity of two documents. Unlike in some of previous approaches (e.g., Mani and Wilson, 2000; Schilder and Habel, 2001),

we are not attempting to establish a chronological order for various actions occurring in the text of a news document. We want to model the temporal similarity in terms of overlap in the temporal references by running pairwise comparisons of documents. The result of comparison characterizes the proportion of overlap in the resolved expressions of two documents. We show that this overlap or *coverage* is higher when two documents discuss the same event than when they are not. Naturally, a TDT system would not make decision based on sheer temporal similarity, but we believe it will provide valuable additional evidence.

We briefly outline some of the previous work by others in Section 2. In Section 3 we describe our approach to extraction and resolving temporal expressions in detail. Section 4 discusses our temporal similarity measure and Section 5 deals with the experiments with a Finnish online news corpus. Section 6 is a conclusion.

2 Related Work

Mani and Wilson (2000) have presented an approach for resolving temporal expressions from news-feed with manually produced and machine-learned rules. They could resolve expressions with explicit ('tomorrow') and positional ('next week') offsets based on the reference time and implicit ('Thursday', 'February') offsets based on reference time and the verb tense. While Mani and Wilson confined their experiments to core expressions and neglected prepositions, Schilder and Habel (2001) introduced prepositions in their finite-state automata using Allen's previous work (Allen, 1983). Both of these approaches make efforts in assigning a temporal tag for events¹ occurring in the news document. Setzer and Gaizauskas (2000) have proposed annotation schemes for such events and temporal information.

Dialogue systems have had similar aspirations to these approaches. For example, Wiebe et al. (1998) presented a machine learning approach for a scheduling dialogue, where two agents, a human and a computer, are trying to agree on meetings via natural language.

Temporal information has been used in detecting events from the text. Smith (2002) employed spatio-temporal attributes in detecting historical events from a digital library. His method employs place-time collocations. Koen and Bender (2000) augmented the news documents with external temporal information and thus anchored the temporal expression occurring in the text into events more familiar to the reader. The temporal references have nurtured the domain of topic detection and tracking in various forms (Allan et al., 1998; Makkonen et al., 2003) as well as more recent area of temporal summarization (Allan et al., 2001).

All of the approaches require a temporal ontology or time-line that provides terms such as *year*, *Monday* and *week* a semantical interpretation. Often time-line simply relies on system date, but temporal algebras have been proposed in order to formalize the time-line and the units of time it contains (Goralwalla et al., 2001; Ning et al., 2002).

3 Extracting Temporal Expressions

In gathering temporal evidence, we need to recognize temporal expressions and to map them onto a timeline. In practise, we map each recognized expression onto a linearly ordered line as a pair of points indicating start and end of the interval.

3.1 Recognition

There are *explicit* temporal expressions, such as 'October 24th 2002', the spotting of which can be conducted with simple patterns. However, good majority of the expressions occurring in the news are either *indexical* (e.g. 'today', 'on Tuesday evening') or *vague* (e.g. '2 months').

¹Here, an event is understood differently from TDT. An event is defined by a verb (e.g., the spokesman *said*) or by event-denoting nouns (e.g., *elections*).

baseterm	engl.
vuosi	year
(vuosi)neljännes	quarter
vuodenaika	season
kuukausi	month
viikko	week
viikonpäivä	weekday
päivä	day, yesterday, ...

Table 1: A list of temporal baseterms.

type	representative terms
{ <i>numeral</i> }	1, 2, 3, ...
{ <i>ordinal</i> }	ensimmäinen, toinen, ...
{ <i>prev</i> }	viime (indeclinable), edellinen
{ <i>ago</i> }	sitten
{ <i>next</i> }	ensi (indeclinable), seuraava
{ <i>after</i> }	kulua+PCP2+PTV, päästä
{ <i>until</i> }	asti, saakka, mennä+INF2+INE
{ <i>onwards</i> }	alkaen, eteenpäin
{ <i>during</i> }	aikana
{ <i>internal</i> }	alku, loppu, ... (beginning, end, ...)

Table 2: The temporal modifiers grouped into sets.

The recognition of these relies on syntactical parsing of the text. If the set of resulting baseforms of given sentence contain *temporal baseterms*, the sentence is passed on to a finite state transducer (similarly to Schilder and Habel, 2001). The temporal baseterms are presented in Table 1.

These keywords can be encountered with certain *temporal modifiers*, such as *next*, *last*, *before*, *ago*, *until*, and together with the keyword they form a *composite expression*, otherwise we have merely *simple expression*. However, in Finnish, these modifiers can be encountered only with certain *cases* of the keyword. We divide the modifiers into 2 types of shift and 2 types of span operations. The shift operations transpose the referent to the left or to the right on the timeline. The span operations open an interval to the left or to the right from the referent on the timeline. For example, the sentence

1. *John called three weeks ago.*

contains a left shift operation that transposes the moment of utterance three weeks to the past. Similarly, the sentence

2. *Until Tuesday it was empty.*

opens an interval of time the end-point of which is some particular Tuesday.

In order to generalize the patterns, we group various prepositions, numerals etc. into the following types presented in Table 2. In the following patterns, curly brackets refer to a set, brackets with upper case contents denote the temporal baseterm and square brackets indicate that the element is not necessary element. If the baseterm is augmented with an asterisk '*', it means that the baseterm refers to a quantity in the shift operation rather than the temporal referent that is being shifted. In the example 1., the moment of utterance is being shifted whereas in 'John called last week' it is baseterm 'week' that is being shifted.

Note: In Finnish, the prepositions are often coded with cases and thus expressions like 'in March' require inessive, 'in last week' adessive, 'in last year' and 'last Saturday' essive, respectively. For example:

1. *Se oli viime marraskuussa.*
it be+PAST last November+INE.
2. *Se oli viime vuonna.*
it be+PAST last year+ESS.
3. *Se oli viime viikolla.*
it be+PAST last week+ADE.

In the following, we present some sample patterns through which different operations can be recognized and corresponding examples.

Left Shift There are several patterns which transpose the baseterm to left on the timeline.

(NOM*) {ago}
'vuosi sitten'
{numeral} (PTV*) {ago}
'kolme viikko sitten'
{prev} (NOM ESS ADE INE)
'viime lokakuussa'
{prev} (GEN) {internal}
'viime kesän lopussa'
{prev} (GEN)[(GEN)] (ESS ADE NOM)
'vuoden vuoden [elokuun] alussa'

Right Shift These relations shift the referent forth.

[{numeral}] (GEN*) {after}
'[kahden] viikon kuluttua'
{next} (NOM ESS ADE INE)
'ensi kuussa'
{next} (GEN) {internal}
'seuraavan vuoden puolivälissä'
{next} (GEN) (ESS ADE NOM)
'ensi viikon maanantaina'

Left Span A keyword can also denote a ending for a period of time.

[{prev next}] (ILL) {until}
'viime sunnuntaihin saakka'
[{prev next}] [(GEN)] (GEN) {internal}{until}
'ensi [vuoden] marraskuun loppuun asti'
(TRL)
'jouluksi'

Right Span A keyword can also denote a beginning for a period of time.

[{prev next}] (ELA) {onwards}
'keväästä alkaen'
[{prev next}] [GEN] (GEN) {internal}{onwards}
'viikon alusta eteenpäin'

Sometimes the patterns are not capable of describing the direction of the span. E.g., the expression 'maanantaista perjantaihin' (from Monday to Friday) refer to the past and just as well to the future regardless of the time of utterance. This applies to good portion of the other expressions as well. Such expressions are resolved via verb tenses.

(ELA) (ILL)
'maanantaista perjantaihin'

3.2 Calendar

In order to formalize the structure of the timeline, we adopt the calendar model Goralwalla et al. (2001) to construct an algebra for temporal expressions.

Definition 1 A global timeline \mathcal{T} is a point structure with precedence relation $<_{\mathcal{T}}$. An interval $[t_i, t_k] \subset \mathcal{T}$, $t_i < t_k$ is a set of instants $\{[t_j, t_j] \mid i \leq j \leq k\}$.

A calendar enables mapping the natural language temporal expressions onto the timeline, i.e., it provides a domain onto which one can map the natural language expressions, explicate their meaning and conduct comparisons, for example. Without a calendar, expressions like 'last April' and 'two weeks ago' could hardly be understood.

Definition 2 A calendar \mathcal{C} is a triplet $\langle \mathcal{T}, \mathcal{G}, \mathcal{F} \rangle$, where \mathcal{T} is the global timeline of \mathcal{C} , \mathcal{G} is the set of granularities $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$. \mathcal{F} is the set of conversion functions between the granularities.

A granularity stands for the valid basic unit of time in the calendar. For example, a Gregorian calendar has a granularities

$$\mathcal{G} = \{G_{year}, G_{month}, G_{week}, G_{day}, \dots\}.$$

A stock exchange related calendar could augment this set with $G_{quarter}$, for example. A granularity is *coarser* if the unit of time it represents is lengthier and *finer* if the unit of time is shorter than that of some other granularity. For example, G_{month} is coarser than G_{day} , and G_{month} is finer than G_{year} . The functions \mathcal{F} convert the quantities in some granularity to some other granularity.

$$\begin{aligned} f_{\mathcal{C}}^{G_1 \rightarrow G_2}(i_1) &\rightarrow N_{G_2}, \\ f_{\mathcal{C}}^{G_2 \rightarrow G_3}(i_1, i_2) &\rightarrow N_{G_3}, \\ &\vdots \\ f_{\mathcal{C}}^{G_n \rightarrow G_{\mathcal{T}}}(i_1, i_2, \dots, i_n) &\rightarrow R_{G_{\mathcal{T}}}, \end{aligned}$$

where $i_j (1 \leq j \leq n)$ is the ordinal of the calendric element of the j^{th} calendric granularity in calendar \mathcal{C} . The result $N_{G_{j-1}}$ is the number of units of G_{j-1} contained by i_1, \dots, i_j . The final outcome, $R_{G_{\mathcal{T}}}$ is the (real) number of the finest units of time on timeline \mathcal{T} . The finest granularity, also *bottom granularity* $G_{\mathcal{T}}$, is the basis for other granularities, and could be, e.g., a day, a second, a nanosecond. Naturally, the values $N_{G_{j-1}}$ have a lower and upper bound depending on the values of i_1, \dots, i_n . For example,

$$\begin{aligned} f_{\mathcal{C}}^{G_m \rightarrow G_d}(2002, 12) &= 31, \\ f_{\mathcal{C}}^{G_m \rightarrow G_d}(2001, 11) &= 30, \\ f_{\mathcal{C}}^{G_m \rightarrow G_d}(2000, 2) &= 29, \\ f_{\mathcal{C}}^{G_m \rightarrow G_d}(1999, 2) &= 28. \end{aligned}$$

The granularity G_{monday} contains every seventh element of G_{day} , and $G_{December}$ every twelfth element of G_{month} . Analogously, $G_{weekend}$ comprises all the Saturday-Sunday pairs.

The conversions between granularities where the coarser cannot be partitioned into integral units of the finer, pose a problem. For example, a calendar month is not evenly divisible into full weeks. Fortunately, for our purposes it is enough that all the temporal expressions are first expressed in terms of the bottom granularity (similarly to Ning et al. (2002)), after which determination of the temporal expression is easier.

3.3 Mapping

Basically, we want to make a mapping from the natural language expressions (name) to discrete linearly ordered space (object). Furthermore, the mapping, i.e., *canonization*, is a bijection, since with the valid temporal expressions name uniquely an interval on the timeline. Sometimes the canonization requires the time of utterance or even syntactic parsing information.

Definition 3 A canonized form is a pair $\langle t_i, t_j \rangle, t_i, t_j \in \mathcal{T}$ of points on the timeline that denotes an interval $[t_i, t_j] \subset \mathcal{T}$ if $t_i < t_j$, or an instant $[t_i, t_i]$ in case $t_i = t_j$.

In the following, we content ourselves with day-level, i.e., $G_{\mathcal{T}} = G_{day}$. Thus all the references to parts of days, e.g., morning, noon, evening, are interpreted as day.

In order to determine the temporal intervals denoted by the baseterms we define a couple of helpful functions:

Definition 4 A function $\pi : \mathcal{G} \times \mathcal{T} \rightarrow \mathcal{T}$, $\pi(G, t_u) = t_i$, returns the starting point t_i of previous element of granularity G with respect to time t_u . Similarly, $\rho : \mathcal{G} \times \mathcal{T} \rightarrow \mathcal{T}$, $\rho(G, t_u) = t_k$ returns the start point t_k of next element in granularity G with respect to t_u .

For example, the element referring to the following Friday with respect to November 4th, 2002, would be $\rho(G_{friday}, t_{20021104}) = t_{20021108}$ while the start of previous winter wtr. to the same day would be $\pi(G_{winter}, t_{20021104}) = t_{20011201}$.

Note that here the i in t_i is not the *ordinal* of the element on the timeline \mathcal{T} but rather the name of it.

Definition 5 A function $\delta : \mathcal{G} \times \mathbb{N} \times \mathcal{T} \rightarrow \mathcal{T}$ shifts the given interval left (δ_L) or right (δ_R) on the timeline \mathcal{T} for n units of granularity G .

In the shifting process, we have to take into account the possibly varying lengths of the elements of G_{months} . For example, 'n months ago' uttered at moment t_i would result in left shift

$$\delta_L(G_{month}, n, t_i) = \delta_L(G_{day}, days, t_i),$$

where

$$days_L = \sum_{k=1}^n f_C^{G_m \rightarrow G_d}(y_i, m(t_i) - k),$$

where the function $m : \mathcal{T} \rightarrow \{1, \dots, 12\}$, $m(t_i) = m_i$ returns the number of months of the given point of time and y_i decreased by one and $m(t_i)$ is set to 12 every time $m(t_i) = k$. For δ_R the amount of days would be resolved with

$$days_R = \sum_{k=0}^{n-1} f_C^{G_m \rightarrow G_d}(y_i, m(t_i) + k),$$

where y_i increased by one and $m(t_i)$ is set to 1 every time $m(t_i) + k > 12$, and

Now, we are ready to define the intervals denoted by the baseterms as presented in Table 3.

Next, we need to provide semantics for the temporal modifiers. As discussed in Section 3.1. Table 4 lists the semantic interpretations of some temporal modifiers. The numerals multiply the shifting granularity. The ordinals usually refer to quarter (1st, 2nd, 3rd, 4th) but they occur in expressions like 'the second weekend in June', as well.

The spans opened by $\{until\}$ and $\{onwards\}$ can result in open-ended spans depending on the tense of the verb. The modifiers contained $\{internal\}$, such as 'alussa' (in the beginning of) stand for a subset of the initial interval. In addition, we simply divide referred interval into three partitions: beginning, middle, and end, and thus expressions like 'in the beginning of/early August 2002' are interpreted as $[t_{20020801}, t_{20020810}]$.

baseterm	start	end
year	$\pi(G_{year}, t_u)$	$\pi(G_{day}, \rho(G_{year}, t_u))$
season	$\pi(G_{season_x}, t_u)$	$\pi(G_{day}, \delta_R(G_{month}, 3, \pi(G_{season_x}, t_u)))$
quarter	$\pi(G_{Q_n}, t_u)$	$\pi(G_{day}, \rho(G_{Q_{n+1}}, \pi(G_{Q_n}, t_u)))$
month	$\pi(G_{month}, t_u)$	$\pi(G_{day}, \rho(G_{month}, t_u))$
week	$\pi(G_{monday}, t_u)$	$\rho(G_{sunday}, t_u)$
day	t_u	t_u

Table 3: A list of temporal baseterm spans with respect to the time of utterance t_u .

Example Explicit expression are canonized regardless of the time of utterance. For example, the expression 'elokuussa 2002' (in August 2002) would yield an interval:

$$\begin{aligned}
& [t_{20020801}, \pi(G_{day}, \rho(G_{month}, t_{20020801}))] \\
= & [t_{20020801}, \pi(G_{day}, t_{20020901})] \\
= & [t_{20020801}, t_{20020831}]
\end{aligned}$$

Example The mapping of the composite expression 'kolme viikkoa sitten' (three weeks ago) uttered on November 4th 2002 transposes given date t_i three weeks backwards:

$$\delta_L(G_{week}, 3, t_i) = \delta_L(G_{day}, days, t_i),$$

where

$$\begin{aligned}
days &= \sum_{k=1}^3 f_C^{G_w \rightarrow G_d}(t_i) \\
&= 7 + 7 + 7 \\
&= 21
\end{aligned}$$

and thus both start and end points of the interval we get

$$\begin{aligned}
\delta_L(G_{week}, 3, t_i) &= \delta_L(G_{day}, 21, t_i) \\
&= t_{20021014}
\end{aligned}$$

Example Finally, the expression 'tään syksynä' ((during) this Autumn) uttered on November 4th 2002 is canonized as follows: The starting point of the resulting interval

$$\begin{aligned}
t_{start} &= \pi(G_{autumn}, t_i) \\
&= t_{20020901}
\end{aligned}$$

modifier	span
{ <i>numeral</i> }	$n \in N$
{ <i>ordinal</i> }	varies
{ <i>prev</i> }	$[\delta_L(G, 1, t_i), \delta_R(G, 1, t_j)]$
{ <i>ago</i> }	$[\delta_L(G, n, t_i), \delta_R(G, n, t_j)]$
{ <i>next</i> }	$[\delta_R(G, 1, t_i), \delta_R(G, 1, t_j)]$
{ <i>after</i> }	$[\delta_R(G, n, t_i), \delta_R(G, n, t_j)]$
{ <i>until</i> }	$[varies, t_j]$
{ <i>onwards</i> }	$[t_i, varies]$
{ <i>during</i> }	$[t_i, t_j]$
{ <i>internal</i> }	$[t_l, t_k] \subseteq [t_i, t_j]$

Table 4: The semantics of the temporal modifiers.

and the ending point of the resulting interval

$$\begin{aligned}
t_{end} &= \pi(G_{day}, \delta_R(G_{month}, 3, \pi(G_{autumn}, t_u))) \\
&= \pi(G_{day}, \delta_R(G_{month}, 3, t_{20020901})) \\
&= \pi(G_{day}, \delta_R(G_{day}, days, t_{20020901})), \\
&= \pi(G_{day}, t_{20021201}), \\
&= t_{20021131},
\end{aligned}$$

where

$$\begin{aligned}
days &= \sum_{k=0}^2 f_C^{G_m - G_d}(y_i, m(t_i) + k), \\
&= 31 + 30 + 31 \\
&= 92.
\end{aligned}$$

4 Temporal Similarity

The documents, or rather their canonized temporal expressions are compared pair-wise. Each start-end pair of one document is compared to each of the start-end pairs of the other. The relations between these intervals fall into following seven categories (Galton, 1995).

- $[t_i, t_j]$ is **before** $[t_k, t_l]$ if $t_j < t_k$,
- $[t_i, t_j]$ **meets** $[t_k, t_l]$ if $t_j = t_k$,
- $[t_i, t_j]$ **overlaps** $[t_k, t_l]$ if $t_i < t_k < t_j < t_l$,
- $[t_i, t_j]$ **begins** $[t_k, t_l]$ if $t_i = t_k \wedge t_j < t_l$,
- $[t_i, t_j]$ **falls within** $[t_k, t_l]$ if $t_i < t_k \wedge t_j < t_l$,
- $[t_i, t_j]$ **finishes** $[t_k, t_l]$ if $t_i < t_k \wedge t_j = t_l$,
- $[t_i, t_j]$ **equals** $[t_k, t_l]$ if $t_i = t_k \wedge t_j = t_l$,

Note that the first six have also the converse relation. In the following, we are not concerned, whether A is **before** B or vice versa.

Motivated by the reliability investigations of Krippendorff (1995) in comparing sets of intervals, we place the two sets on a cross-tabulation as illustrated in Figure 1. The diagonal represents the synchronous points of the two time-axes, and the shaded areas correspond to the intersections between two intervals $A = \{A_1, A_2, A_3\}$ and $B = \{B_1, B_2, B_3, B_4\}$.

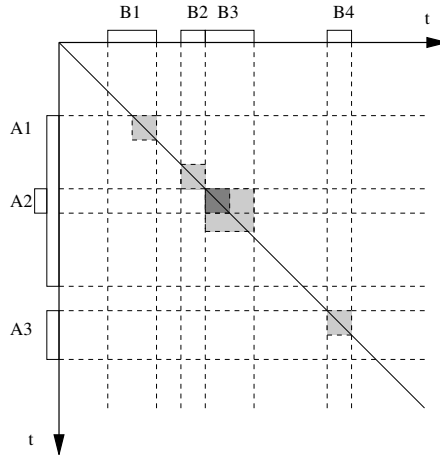


Figure 1: A cross-tabulation of two sets of intervals A and B .

Should the two sets be identical, the shaded areas would be of size $A_i \times B_i$ for each i . In such case, all of the intervals of A would be covered with and an interval of B . We do not wish to attribute each kind of relation a uniform emphasis as to similarity, since we would value more

knowing a day falls within a weekend or a week than falling within a year’s interval. Therefore, we propose a simple weight function $\mu : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$ such that

$$\mu([t_i, t_j], [t_k, t_l]) = \frac{2 \Delta([t_i, t_j] \cap [t_k, t_l])}{\Delta(t_i, t_j) + \Delta(t_k, t_l)},$$

where $\Delta : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$, $\Delta(t_i, t_i) = 1$ is the duration (in days) of the given interval. The weight function results in 1 if the expressions are an exact match and 0 if the expressions are distinct. All of the relations presented above can be expressed with the duration of the intersection, because we do not make the distinction between **begins**, **falls within** and **finishes**. Rather, they are just cases of having a non-empty intersection.

As to the example of Figure 1, the intersections $A_3 \cap B_4$ and $A_2 \cap B_3$ would result in higher μ -value than the any of the intersections $A_1 \cap B_1$, $A_1 \cap B_3$, and $A_1 \cap B_2$, because the sizes of the intersections $A_3 \cap B_4$ and $A_2 \cap B_3$ are closer to the sums $|A_3| + |B_4|$ and $|A_2| + |B_3|$.

In practice, the pair-wise μ -weights are calculated in what we call a *cover matrix* illustrated in Table 4. The coverage of an interval $T_{i,j}$ is calculated by choosing the maximum $v_{i,j}$ of the of the weights for that term. If an interval $T_{1,i}$ is covered with an interval $T_{2,j}$ of equal weight, the maximum value is $v_{1,i} = 1$. On the contrary, if it is not covered all, the maximum value yields $v_{1,j} = 0$. In cases of inclusion the cover varies in $(0, 1)$ depending on the sizes of the intervals.

	$T_{2,1}$...	$T_{2,m}$	max
$T_{1,1}$	$\mu(T_{1,1}, T_{2,1})$...	$\mu(T_{1,1}, T_{2,m})$	$v_{1,1}$
\vdots	\vdots		\vdots	
$T_{1,n}$	$\mu(T_{1,n}, T_{2,1})$...	$\mu(T_{1,n}, T_{2,m})$	$v_{1,n}$
max	$v_{2,1}$		$v_{2,m}$	

Table 5: A cover matrix.

The total coverage of the two sets of intervals is the sum of all the maximum values divided by the number of interval, i.e.,

$$coverage(T_1, T_2) = \frac{\sum_{i=1}^n v_{1,i} + \sum_{j=1}^m v_{2,j}}{m + n}$$

5 Experiments

Our corpus consists of 3958 Finnish online news documents from a single Internet source dating from April 1 to June 30 2001. Each document had about 100 words and 2.3 temporal expressions. An event is *something that happens at some specific time and place* (Yang et al., 1999). We have manually assigned the documents to a total of 77 events of varying size and type. A total of 398 documents was labeled with some event that contained 2 or more documents. On the average, each event was discussed in 5.16 documents.

5.1 Extraction

We experimented the temporal expression recognition and canonization by manually evaluating the output of the extractor with a modest subset of the corpus of 134 documents. There were 1194 sentences in all with 322 temporal expressions 95 of which were composite and 227 were simple.

The recognition of the patterns performed quite decently. The canonization did not behave equally well. There were problems especially in determining the direction of the shift or span. In sentences such as 'The incident is a result of Friday’s bloodshed.' or 'The election was supposed to take place on 23rd of May.' the tense of the verb misleads to opposite direction. The resolving would require external knowledge.

type	precision	recall
simple	0.98	0.91
composite	0.95	0.81

Table 6: The results from temporal expression extraction.

In addition, improvement would require thorough enumeration of the variants of words: 'kolme' (three), kolmisen (about three, *threeish*). Furthermore, there is a plethora of expressions that enforce vagueness to the expression the formalization of which is always ambiguous.

5.2 Event Identification

We calculated the average distribution of the temporal relations of the intervals. The results in Table 7 supports the use of temporal evidence in identifying events. Although the average percentage of distinct time-intervals is slightly higher when comparing the documents discussing the same event, the amount of exact matches is considerably higher within same event.

relation	same event	
	yes	no
before	89.28	89.00
meets	0.01	0.02
overlaps	0.05	0.13
begins	0.82	0.32
falls within	5.44	9.49
finishes	0.54	0.20
exact	3.84	0.83

Table 7: The average distribution of temporal relations.

The granularities of Gregorian calendar we used do not meet or overlap each other very often, while the various sorts of inclusion are more frequent. Usually, the overlapping intervals are a week and a month.

We also performed temporal comparisons using the coverage-function and the sum of μ -values. The values within the same event are twice the values within different event, as shown in Table 8.

	same event	
	yes	no
$\sum \mu$	0.0049	0.0025
coverage	0.0072	0.0034

Table 8: The averages of the sum of μ -values and the coverage.

6 Discussion

We have presented an approach to extract temporal expressions and to formalize their meaning for Finnish, though the general approach would easily convert to other languages. In the process the natural language expressions to a general timeline, we employed finite state transducers and a calendar framework that we augmented with shift and span operations. The performance of the extraction of the temporal evidence was encouraging.

We also presented an method for comparing extracted temporal evidence, and employed it in identifying events from the news feed. Although detecting events with mere temporal evidence would hardly be feasible, it can be used to augment the information deriving from comparison between terms, places, names etc. as it has be reported in (Makkonen et al., 2003).

There is a multitude of exceptions and special cases not covered in the patterns. We confined our made efforts in providing a simple interpretations for the most typical portion. Another area for improvement would be incorporating the delicate difference between *utterance time* and *reference time* (see e.g., Allen, 1983), where some of the expressions would not be evaluated with respect to time of utterance but possibly the previous expression.

References

- Allan, J. (Ed.) (2002). *Topic Detection and Tracking – Event-based Information Organization*. Kluwer Academic Publishers.
- Allan, J., J. Carbonell, G. Doddington, J. Yamron, and Y. Yang (1998). Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*.
- Allan, J., R. Gupta, and V. Khandelal (2001). Temporal summaries of news topics. In *Proc. ACM SIGIR*, pp. 10–18.
- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of ACM* 26(11), 832–843.
- Galton, A. (1995). Time and change for AI. In M. Gabbay, C. J. Hogger, and J. A. Robinson (Eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 4, Epistemic and Temporal Reasoning*, pp. 175–240. Oxford University Press.
- Goralwalla, I. A., Y. Leontiev, M. T. Özsu, D. Szafron, and C. Combi (2001). Temporal granularity: Completing the puzzle. *Journal of Intelligent Information Systems* 16(1), 41–63.
- Koen, D. B. and W. Bender (2000). Time frames: Temporal augmentation of the news. *IBM systems journal* 39(3&4), 597–616.
- Krippendorff, K. (1995). On the reliability of unitizing continuous data. In P. V. Marsden (Ed.), *Sociological Methodology*, Chapter 2, pp. 47–76. Blackwell.
- Makkonen, J., H. Ahonen-Myka, and M. Salmenkivi (2003). Topic detection and tracking with spatio-temporal evidence. In *Proc. 25th European Conference on Information Retrieval Research (ECIR 2003)*, Pisa, Italy, pp. 251–265.
- Mani, I. and G. Wilson (2000). Robust temporal processing of news. In *Proc. Association for Computational Linguistics (ACL)*, pp. 69–76.
- Ning, P., X. S. Wang, and S. Jajodia (2002). An algebraic representation of calendars. *Annals of Mathematics and Artificial Intelligence* 36(1-2), 5–38.
- Schilder, F. and C. Habel (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proc. ACL-2001 Workshop on Temporal and Spatial Information Processing*, pp. 65–72.
- Setzer, A. (2001). *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Ph. D. thesis, University of Sheffield, UK.
- Setzer, A. and R. Gaizauskas (2000). Building a temporally annotated corpus for information extraction. In *Proc. 2nd Intl. Conference on Language Resources and Evaluation (LREC) workshop: Information Extraction Meets Corpus Linguistics*.

- Smith, D. A. (2002). Detecting events with date and place information in unstructured text. In *Proc. 2nd Joint Conference on Digital Libraries (JDCL'02)*, pp. 191–196.
- Wiebe, J., T. O'Hara, K. McKeever, and T. Öhrström-Sandgren (1998). An empirical approach to temporal reference resolution. *Journal of Artificial Intelligence Research* 9, 247–293.
- Yang, Y., J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval* 14(4), 32 – 43.